PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 :
H03K 19/00, G06F 9/455, 17/50, G06G 7/48

A1

(11) International Publication Number: WO 00/44094

(43) International Publication Date: 27 July 2000 (27.07.00)

(21) International Application Number: PCT/US00/01360

(22) International Filing Date: 20 January 2000 (20.01.00)

(30) Priority Data:
60/116,714 21 January 1999 (21.01.99) US

(71) Applicant *(for all designated States except US)*: UNIVERSITY OF SOUTH CAROLINA [US/US]; Osborne Administration Building, Suite 109, Columbia, SC 29208 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: TOUR, James, M. [US/US]; 3700 Linbrook Drive, Bellaire, TX 77401 (US). REED, Mark, A. [US/US]; 129 Guinea Road, Monroe, CT 06468 (US). SEMINARIO, Jorge, M. [US/US]; 1124 Autumn Circle, Columbia, SC 29206 (US). ALLARA, David, L. [US/US]; 2517 Sleepy Hollow Drive, State College, PA 16801 (US). WEISS, Paul, S. [US/US]; 545 Orlando Avenue, State College, PA 16801 (US).

(74) Agents: HARDAWAY, John, B., III et al.; Nexsen Pruet Jacobs & Pollard, LLP, Post Office Drawer 10648, Greenville, SC 29603-0648 (US).

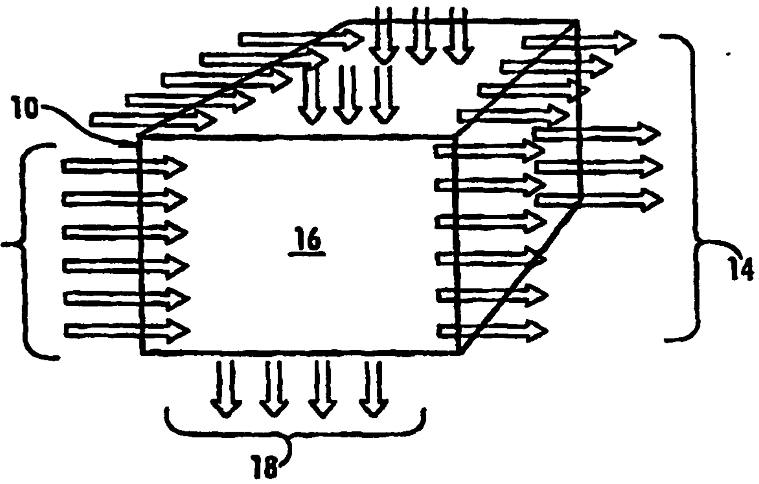
(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: MOLECULAR COMPUTER

(57) Abstract

A molecular computer is formed by establishing arrays of spaced-apart input (12) and output pins (14) on opposing sides of a (10), injecting moleware (16) in solution into the containment and then allowing the moleware to bridge the input and output pins. Moleware includes molecular alligator clip-bearing 2-, 3-, and molecular 4-, or multi-terminal wires (30, 36, 40, respectively), carbon nanotube wires (80), molecular resonant tunneling diodes (32, 38), molecular switches (90), molecular controllers (42) that can be modulated via external electrical or magnetic fields, massive interconnect stations (44) based on single nanometer-sized particles, and dynamic and static random access memory (DRAM and SRAM) components composed of molecular controller/nanoparticle or fullerene hybrids (46). The current-voltage characteristics that result from the bridging between input and output arrays can be ascertained using another computer to identify the bundles of inputs and corresponding outputs that provide a truth table for the specific functions of the computer.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania
AM	Armenia
AT	Austria
AU	Australia
AZ	Azerbaijan
BA	Bosnia and Herzegovina
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CU	Cuba
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GH	Ghana
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IL	Israel
IS	Iceland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgyzstan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakstan
LC	Saint Lucia
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LS	Lesotho
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
MK	The former Yugoslav Republic of Macedonia
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TM	Turkmenistan
TR	Turkey
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam
YU	Yugoslavia
ZW	Zimbabwe

MOLECULAR COMPUTER

1. Priority Statement:

We claim the benefit of the priority date of US Provisional Patent application 60/116,714, filed January 21,1999, which is incorporated herein by reference.

2. Field of the Invention:

The present invention relates generally to computer architecture, and, more specifically to a computer that uses molecules as functional units (e.g. logic gates and memory cells).

3. Background of the Invention

The investment by society in computer technology has been astonishing in both its rate of increase and its extent. In less than three decades, the personal computer, for example, has gone from experimental prototype to being an essential tool of business. Demands for computers with greater and greater capabilities to perform more and more tasks continue unabated. Heretofore, better computers have resulted from increased miniaturization, among other improvements.

Presently, however, we do not have a viable technology for our near future computer needs. Although Moore's law (an accurate empirical law at this time) predicts the doubling of computer power every 18 months, this trend cannot continue. Digital computers are presently based on silicon technology. More precisely, very large scale integration (VLSI) is a lithographic technology, and although silicon is indeed quite important, Moore's Law is essentially lithography driven. The law of diminishing returns will eventually conquer Moore's Law, perhaps by 2005, when the cost of an integrated-circuit fabrication factory will become exorbitant and spell the demise of the growth of VSLI-based computer systems.

Other than standard logic and memory tasks, a technique one might choose for an ideal post-VLSI computer system would be to utilize a non-lithographic approach to construction, e.g., a directed self-assembly, so there would either be a vast redundancy or the system should have fault tolerance through dynamic fault reduction. Another property for such a computer would be an ability to be self-reconfigurable, that is, it should be able to dynamically reconfigure interconnects in response to inputs, or else it will also fall victim to the incessant demands for interconnects. Finally, this post-VLSI computer would be innovative: able to reconfigure architectures not previously experienced.

There remains a need, therefore, for a new computer technology that provides the advantages of self-assembling construction and greater processing capability that can be reconfigured readily.

SUMMARY OF THE INVENTION

The present invention is a device having a multiplicity of input pins and a multiplicity of output pins on its external surface. The interior is composed of a self-assembled array of specifically selected and adapted molecules, called "moleware", bridging the inputs and outputs. Initially, this present computer is in a blank state; that is, nothing is known about the electrical-signal-transferring relationship between the input and output pins. A voltage is applied to each of a series of bundles of inputs. The outputs are searched, also in bundles, to determine what outputs have signals running through them. Using a computer to do the searching, sets of bundles of inputs and outputs will eventually be identified that can be used to constitute a truth or partial truth table for the computer. Electrical or magnetic fields can also be applied across the container to increase configuration possibilities.

The present invention, in a preferred embodiment, may take the form, for example, of a one cubic centimeter box that contains 1000 metallic inputs (m), 1000 outputs (n), and 100 different learning inputs (l). The box will contain an intelligently self-assembled array of 10^{20} pre-designed active and passive components ("moleware") including molecular alligator clip-bearing 2-, 3-, and molecular 4-terminal wires, molecular resonant tunneling diodes, molecular switches, molecular controllers that can be controlled via external electrical or magnetic fields, massive interconnect stations based on single nanometer-sized particles, dynamic or static random access memory (DRAM or SRAM) components composed of molecular controller/nanoparticle or fullerene hybrids. Moreover, the present molecular computer, once formed, can be further modified by new interconnect routes via electrochemically induced cross linking of the nanoparticles or by "burning out" components through large induced fields, analogous to the operation of a field programmable gate array (FPGA). Arrays of these molecular computers can be coupled using standard interconnect methods to form massive molecular computer arrays.

The present molecular computer has a number of advantages:

It can be teachable. The system can be trained by forcing the correct output for a given input by varying the operational inputs until consistency is achieved. The output as a function of input can be autoverified.

It can be reconfigurable. Either by successive retraining or by burning out specific functions, the system can be changed.

It can possess logic. The truth table for the molecular computer would not be known a priori, but would be determinable once it has been prepared or on the fly as self-assembly takes place

It can possess memory. Semiconductor nanoparticles such as CdS or CdSe coated with molecular control elements, or C_{60} surrounded by controller elements, or even nanometer-scale metallic particles sufficiently small so as to exhibit sizeable Coulomb blockade, will be preformed and then permitted or induced to assemble themselves as part of the network of the present molecular computer. In the situation where an electromagnetic field is applied, the controllers will open thereby permitting the semiconductive or metallic particle to store a charge, hence acting as a DRAM or SRAM component.

It will have an intrinsic extremely high fault tolerance because of the ability to use their terminals (input, outputs, and learning inputs) as rewiring control leads. Therefore, this embodiment takes advantage of the inmense number of functions that can be make with their terminals. For example, for binary operation, 1000 inputs yield 2^{1000} posible input combinations. Therefore with 1000 possible outputs yield a total of $10^{32256000..........0000}$ where the number of digits in the exponent is 304.

The system has dynamic fault reduction capabilities. Unlike VLSI, where the interconnect structure is rigid, the present molecular computer can be reprogrammed to eliminate undesired and inoperative fragments since the interconnect topology can be rewired at any time by applying sufficiently strong fields to regions of the molecular computer where reprogramming is desired.

The system lends itself to expansion. Arrays of molecular computers according to the preferred embodiment described above can be formed using standard interconnect methods to form massive computer arrays.

It is also possible to program the present computer in other than a binary digital mode, such as for example, an analog or multilevel (semi-analog) system. Perhaps programming becomes more formidable; however, data storage increases significantly and makes the programming challenges acceptable for certain applications.

The design of a molecular computer according to a preferred embodiment of the present invention takes into consideration that a very large percent of the fabricated moleware components may be defective. Hence, the testing will be more extensive and done in multiple stages in the fabrication and assembly. Molecular computers made according to the present process will be tested individually and in place. Testing is performed preferably with a high-speed computer or supercomputer initially or, eventually, with the present molecular computer, due to the computational demands for rapidly sorting out the relationship between input and output pins.

An important feature of the present invention is that, unlike other approaches to making computers from molecules, specific placement of nanometer-sized molecular elements is not needed. Moleware is simply added in bulk in a dilute solution and given a chance to form links or "bridges" between inputs and outputs. The excess solvent is removed leaving the moleware components after self-assembly. Additionally, the moleware can be inserted using an evaporation process wherein the molecules are evaporated into the container wherein they self-assemble. The moleware is selected and adapted to form stable links that will remain in place during normal operating conditions. The molecular computer has then produced inputs and outputs with the necessary relationships; it remains the task of another computer to find which groups of input pins have what relationship to which output pins. Once the input and output pin relationships have been identified, the computer can then be used just as surely as if the moleware had been put carefully into position.

Another important feature of the present invention is that, because the present moleware is capable of transferring voltages comparable to current integrated circuits, the adaptation required to integrate molecular computers into current electronic devices is straightforward.

Still another important feature of the present invention is that it is highly fault tolerant. Because there are vast numbers of molecular paths that are possible between inputs and outputs, it will not matter if a very large percentage of them do not work. Enough of them will work to permit a truth table to be found.

Another important feature of the present invention is the ability to reconfigure it – to increase, change or break connections in the bridging between inputs and outputs – simply by subjecting it to electrical or magnetic fields long enough and strong enough to cause the moleware to realign itself. These fields can be sufficiently local so as not to affect all moleware in the container.

Still another feature of the present invention is that, although each computer will be unique, all can be "trained" to perform the same functions once the necessary relationships between inputs and outputs are found.

Other features and their advantages will be apparent to those skilled in the art of nanotechnology and molecular computers from a careful reading of the Detailed Description of Preferred Embodiments accompanied by the following drawings.

DETAILED DESCRIPTION OF THE DRAWINGS

In the drawings:

Fig. 1A illustrates in schematic form a molecular computer according to a preferred embodiment of the present invention;

Fig. 1B illustrates four species of moleware for use inside the containment of the molecular computer;

Fig. 2A illustrates in schematic form the inputs/outputs for a molecular computer according to a preferred embodiment of the present invention;

Fig. 2B illustrates a detailed perspective view of the inputs/outputs;

Fig. 3 illustrates a two-dimensional projection of a detail of a cross sectional view of the input/output arrangement, according to a preferred embodiment of the present invention;

Fig. 4 illustrates in chemical symbols and schematic form the formation of molecular alligator clips on the end of molecular wires and its attachment by sulfur bonding to metal surfaces, according to a preferred embodiment of the present invention;

Fig. 5 illustrates two additional molecular alligator clips, according to alternative preferred embodiments of the present invention;

Fig. 6 illustrates in chemical symbols a molecular wire with alligator clips according to a preferred embodiment of the present invention;

Fig. 7 illustrates a controller molecule in two orientations, in the "off" position on the left and in the conductive position on the right, according to a preferred embodiment of the present invention;

Fig. 8 illustrates the connection to a gold nanoparticle connected to an electrode via a bi-functional molecular wire, according to a preferred embodiment of the present invention;

Fig. 9 illustrates a nanoparticle with connecting molecules attached to activate its surface and facilitate interconnection with other moleware, according to a preferred embodiment of the present invention;

Fig. 10 illustrates a molecular dynamic random access memory composed of a molecular controller and molecular capacitor, according to a preferred embodiment of the present invention; and

Fig. 11 illustrates a preferred insulator nanoparticle;

Fig. 12 illustrates modular preassembly of the moleware according to a preferred embodiment of the present invention;

Fig. 13 illustrates two-, three- and four-terminal molecular moleware species, according to a preferred embodiment of the present invention; and

Fig. 14 illustrates a two-dimensional array for a molecular computer according to an example of an embodiment of the present invention.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

Figs. 1A and 1B illustrates the configuration of a molecular computer according to a preferred embodiment of the present invention. The computer comprises a containment, illustrated here in the form of a container 10 with input 12 and output ports 14 (I/O ports) and moleware 16 assembled inside the containment. The shape of container 10 is not critical as long as it provides acces to input ports 12 and output ports 14. Current or a field gradient 18 may be applied across container 10. Additionally, a two-dimensional chip-based scheme could be used wherein moleware branches laterally across I/O locations. The term "moleware" is intended to refer to nanometer-scale objects (less than 1000 nanometers in diameter) that serve as part of a computing device; i.e., nanometer-scale computer "hardware" (see Fig. 1B). In general, moleware consists of molecules, clusters of molecules, nanoclusters of atoms, and combinations of these, selected for particular characteristics and adapted to interconnect in specific ways. The moleware is preassembled in standard chemical lab ware containers, preferably in dilute solutions, and is injected into the containment in a controlled sequence that allows the moleware to bridge the input and output ports. The supporting matrix for the input and output pins could be a rectilinear two or three dimensional grid fabricated on substrates made, for example, of glass or silicon. However a substrate-less approach is also possible. The three-dimensional grid can be formed using multilevel fabrication and micro-electromechanical release structures, and moleware can be deposited by evaporation.

Nominally, the present computer can be made, for example, of a one cubic centimeter containment with 1000 input and 1000 output metalic leads. The exact shape of the computer and the number of leads is not critical. All leads, for example, can be on the same side of the containment. The moleware will comprise an intelligently assembled array of about 10^{20} pre-designated active and passive components, called herein "moleware", including molecular alligator clip-bearing 2-terminal wires 30; a 2-terminal wire with a molecular resonant tunneling diode 32; 2-terminal wires with an incorporated quantum well 34; 3- terminal wires 36; 3-terminal wires 38 with molecular resonant tunneling diodes; 4-terminal wires with molecular resonant tunneling diodes 40 (with-diode-version only shown), molecular controllers 42 that can be modulated via the external peripheral fields, massive interconnect stations based on single gold or silver nanoparticles 44, DRAM components 46 composed of molecular controller/nanoparticle or fullerene hybrids in an inert atmosphere, preferably argon or nitrogen. Moreover, the present molecular computer, once formed, can be further modified by new interconnect routes via electrochemically-induced crosslinking of the nanoparticles or by "burning out" components through large induced fields, analogous to the operation of a field programmable gate array (FPGA). Adding, deleting and changing the bridges from input to output pins alters the selection of relationships between the input and output pins.

Initially the system is in a blank state, with random interconnects yielding generally meaningless output for a given input. The system is then trained, that is, taught new logic functions by forcing the correct output n for a given input m (i.e., an n(m) function) by varying the inputs l until stability is achieved. The goal of self-reconfigurability implies that one sacrifices control over detailed reconfiguring: it is not necessary to know the exact interconnect structure in order to give the correct relationship between input and output. The system then functions as a molecular computer. Furthermore, it does not necessarily have to operate in a binary mode. Use of hexadecimal or analog modes, for example, would vastly increase the information storage density at the expense of increased programming demands.

During the assembly, the computer will be connected using the I/O leads. After the assembly process, the computer will be interfaced via the metallic input and output pins of the molecular computer via normal lithographic wiring. For each logical component, an I/O determination routine will be executed up to the point that a usable truth table is obtained for a desired operation. To decrease the number of individual single bit I/O combinations to be tested, I/O tests can be bundled, thereby decreasing the number of I/O testing combinations while increasing the chances of overcoming single fault locations. To determine single fault locations, subcircuits can be tested in aggregate.

More complex, analog responses can be measured as well (a hybrid on-chip multiplexer can be incorporated as necessary) to allow approximation of the numerical parameters associated with the signal thresholds and response curves and to develop and refine software models of the performance of working moleware. This approach allows tuning of the manufacture of the moleware to increase yields of working components.

The programming of the molecular computer takes into account its architecture, and in particular (i) its very large number of components, (ii) its variable assembly process, and (iii) the incorporation of some components that are only partially functional.

The large number of components dictates that a modular programming approach wherein key functional modules are determined and separated for execution by the molecular computer. These modules are trained and tested. The variable assembly of the molecular computer permits a corresponding flexibility in programming. For example, if one wanted relatively more memory capability, more semiconductor nanoparticle/controller (DRAM) elements would be attached during the assembly process. Second, as with the FPGA, specific elements could be hard-wired into the final system by a burning out approach, just as is currently done with macroscopic field programmable arrays. Thirdly, new interconnect routes via electrochemically induced crosslinking of the nanoparticles can be achieved by adding, for example, pyrrole to form new wire interconnects between the particles by application of a potential for polypyrrole growth.

While the system will also be software programmable once a truth table pattern has been defined, incorporation of components wherein a portion of them are only partially functional requires that the programming of the present molecular computer to provide dynamic faults. The software needs the capability to detect faulty components on an on-going basis, and the programming will need to allow faulty components to be bypassed by other components. There is a large body of literature in the error-resilient programming techniques that may be employed. To facilitate the efficient detection and repair of faults, our software architecture for fault resiliency may be designed with a hierarchical structure and employ methods for re-assigning computations to reliable components.

The present molecular computer overcomes shortcomings of VLSI computers by prefabricating moleware, which is defined as molecular components such as RTDs and DRAMs, as well as molecular level interconnects based on molecular wires and junctions. Prefabrication is achieved by intelligent self-assembly strategies. In this context, "intelligent" means using pre-formed molecular components and knowing the order of additions for each component -- typically, molecular wires first, then nanoparticles, then more molecular wires including wires with barriers. In that way, if one desired, a similar molecular computer can be constructed using the same sequence of moleware addition. However, just as with biological brains, no two molecular computers would be identical.

Figs. 2A, 2B and 3 illustrate the containment and pin array. The container 50 for the molecular computer may either contain or be directly fabricated from 2D rectilinear I/O grids 52 on a substrate and be approximately 1 centimeter on a side. Initially, there would be ports 54 for injection of organic solutions of moleware. There would preferably be 1000 pins 60 on each face for input and output. Any combination of the pins could be used for inputs 12 and outputs 14. The I/O routes could be interchanged as designed for a specific function. The interior portions 62 of the pins would be the binding sites to the initial moleware components in the intelligent self-assembly process, or be multi-branched electrodes (trees) that extend into the molecular computer. To provide more states, a multi-branched 3D electrode could have alternating and possibly aperiodic regions, that may be exposed, to form complex nucleation topologies. The initial port 54 could be sealed after assembly is complete. Various electrode issues (metal, spacing, etc.) are within the capabilities of those with ordinary skill in the art without undue experimentation and can be determined on a microfabricated 2D chip with the candidate molecular computer moleware.

Along the periphery or through the container, field gradients can be introduced to control the DRAM segments. These are applied by macroscopic leads running through the interior of the box when higher gradients are to be used. Initially these are nested, insulated multi-branched 3D electrodes that provide the scaffolding for the moleware.

Figs. 4 through 11 illustrate alligator clips, molecular wires and their attachment to nanoparticles and other moleware used with the present invention. One type of molecular wire is illustrated in Fig. 4, namely, a wire 70 with an isocyanate terminus that will interact with the empty $d_z{}^2$ orbital on vanadium 72 and the filled surface bands. Fig. 5 illustrates two other metal complexes, a tetrahedral complex 74 where "M" can be Vanadium, Niobium, or Tantalum, and a square pyramidal complex 76 where "M" can be Molybdenum or Tungsten, for use as a terminus. Fig. 6 illustrates a molecular wire 80 with "alligator clips" 82 (where "Z" can be, for example, SH, $SiCl_3$, $NV(SH)_3$) on each end with solubilizing groups 84 (where "R" can be, for example, $C_{12}H_{25}$) attached along the length of wire 80. "Alligator clip" means a device – here a molecule – that attaches one component to another, in an analogy to macroscopic spring biased "alligator" clips used in more traditional electronics to attach wires to terminals.

The molecular wires are constructed from oligo(phenylene ethynylene)s bearing thiol termini for alligator clips or more conductive planar oligomer segments, or even carbon or related nanotube structures. The wires can be varied in length, functionality, number of termini, etc. Ab initio calculations and one-dimensional simulations of these systems can be performed to determine their characteristics as molecular wires and termini.

The molecules that form the control moleware (Fig. 1B) are those that permit more or less current to flow through them, as modulated by a third electrode field, that will be produced by the field arrays in specific regions as needed. Gold, platinum and silver, perhaps in combination, provide excellent stations for massive interconnection of the assembly moleware, thereby making the contents more conductive and providing for network-like operation. Self-assembly techniques using bifunctional molecular wires can be used to cap the particles with active surface chemical groups that will electrically connect the particles to the device electrodes and to "bare" metal particles of selected sizes, thus forming controlled-geometry, conductive networks. The use of controller molecules in the capping step would add a switching capability to the network lines when external fields are applied.

Using semiconductor nanoparticles such as CdS and CdSe or even fullerenes, and decorating their surfaces with the controller molecules (Fig. 8) would permit the nanoparticles or fullerenes to act as switchable capacitors: charging in the presence of an applied field and retaining the charge once the field is diminished.

Insulating oxides such as Al_2O_3, TiO_2, and SiO_2, will be surface capped with chemical groups for the selective interconnection to conductive metal particles (Fig. 9). These oxide particles thus act as high impedance resistors in the circuit. Control of the fractions of insulating particles in the self-assembling mixture will thus allow control of the total number of circuit connections up to the ultimate limit of the percolation threshold. Approaching this limit will impart exceedingly high sensitivity to the current-voltage characteristics across the individual I/O ports in the device.

All of the moleware: wires, junctions, DRAMs, etc., are synthesized prior to the intelligent self-assembly within the container, as seen in Fig. 12. One only needs to add them and permit the assembly as needed. If more memory elements are desired, for example, one simply adds more DRAMs. By varying the chemical nature of the surfaces of each type of nanoparticle through molecular capping techniques, the relative interconnection affinities of the molecular and nanoparticle components can be controlled thus providing a rational means to tune the final structure and functional states of the molecular computer. This strategy, together with the control of the relative amounts of components, their addition sequence and intermittent real-time feedback driven pulsing of the device I/O ports, form the key to the overall intelligent self-assembly approach. It will be clear that the order of addition of each species of moleware, the amount, the delay time before a second and a third species is added, will affect the development of the nature of the structure bridging the inputs and outputs. Although the bridges so formed will vary, the bulk of them will reflect the rationality of the sequence and a greater percentage will be usable than if all the moleware species were added at the same time. Field effect structures (molecules that change their orbital structures based on electric fields) could also be considered as moleware components.

The alligator clips for the moleware of the present microcomputer may be fabricated using metallic connection atoms in the series sulfur to selenium to tellurium, and others such as isonitrile, carboxylate (for adhesion to the native oxides of Titanium and Aluminum) and $SiCl_3$ (for adhesion to the native oxide on silicon). Alternatively, organo-transition metal species can be made using V, Nb, Ta, Mo, and W, which have an empty low-lying orbital projected toward the metallic surface which should respond as a gradient for the electron transport. Isocyanides are also useful in that they have a two-fold decrease in barrier height when attached to palladium with respect to thiols on gold. Thiols themselves bind well to Au, Ag, GaAs, PtIr and numerous other metallic surfaces. This would project the transition metal atom's empty d_z^2 orbital directly into the metal surface for interaction with the filled metal band (Fig. 4). This is a unique approach to make a system that is analogous to a gradient of charge in a solid state system wherein the gradient is used to solve the impedance mismatch problems.

As seen in Fig. 6, molecules will be functionalized with alligator clips in order to make direct chemical contacts with the I/O electrodes and nanocomponents, viz., nanoparticles, to which the molecules will be attached in functional chains. The chemical nature of the alligator clip groups will be optimized for chemical bonding selectivity to desired surfaces, bond stability and favorable electrical junction properties. The groups will be chosen from a set of precursors which will provide the following bonding units as needed: RS-, RSe-, RTe-, RNC-, RCO_2-, where R = the molecules to which the group is attached. This set allows good flexibility in terms of the possible groups but other groups are possible also such as RPO_3H_2, $RSiO_X$, and RNH_2, depending on the type of electrode or nanoparticle surface desired. The variety of groups goes well beyond the typical S-based ones previously used since these newer varieties includes attachments to insulator surfaces, in addition to the typical case of just Au electrodes used in the past. This extension requires no more than state of the art surface chemistry.

Molecular wires can be fashioned of oligomers, fully soluble, of precise length and with alligator clip connection points. Suitable materials for molecular wires include carbon nanotubes, which are highly conductive, planar polyphenylenes, polypyridines, polypyrazines and polythiophenes.

Fig. 7 illustrates a controller molecule 90 connecting two nanoparticles 92, illustrated in its two states ("on" and "off"). Controller molecules 90 are switched by an electric field. A donor/acceptor component on adjacent rings in an applied field stabilizes the zwitterionic form making it more planar. In this manner, the starting orientation to the applied field is irrelevant so no polling of the system during formation is required. This molecule would be analogous to a bi-polar switch in the "off" position, meaning negligible current would pass though the molecular chain. Once the field is applied, however, the central molecular element would switch to a more planar state with a smaller bandgap, and hence a more conducive configuration to enable current to flow.

The terms planar and perpendicular are not intended in the literal sense, only that the initial structure has a higher degree of twist than the latter structure. Hence electron transport will be more facile in the second structure. This is precisely what is needed for an electronic system. Note that DRAM systems generally are refreshed every 10^{-5} seconds, therefore, even slight electronic retardations are sufficient. The initial biphenylnitroamine moiety would reside in a less planar state than the zwitterionic structure. However, a perpendicular electric field component would stabilize the zwitterionic state, thereby making that form a greater contributor to the resonance forms. Once in that form, electronic passage should be more facile. Even if there were multiple molecules attached to the nanoparticle, as long as one became more planar in the applied field, electrons could be trapped on the nanoparticle and the field then diminished for information storage.

Gold, platinum and silver nanoparticles 100 provide excellent stations for massive interconnection of the assembly moleware, thereby making the contents more conductive and providing for network-like operation. Self-assembly techniques using bifunctional molecular wires 102, best illustrated in Fig. 8, can be used to cap the nanoparticle 100 with active surface chemical groups 104 that will electrically connect the nanoparticles to the device's electrodes 106 and to "bare" metal particles of selected sizes added to the interior of the containment, thus forming controlled geometry conducting networks. The use of controller molecules in the capping step would add a switching capability to the network lines when external fields are applied. This requires self-assembling, nanometer-scale components into strings or networks in which the various components are interconnected in various ways. The basic components consist of multiple alligator clip molecules, including both wires and controllers, and nanoparticles made from metals, semiconductors and insulators. The simplest structure is a nanoparticle, such as would be formed from a semiconductor, attached to an electrode by a bifunctional molecular wire. When a field is applied between the base electrode and a nearby electrode, this structure will act as a capacitor as charge is pulled into (or drawn out of) the semiconductor particle. A more complicated linear string would contain a field activated controller molecule and a semiconductor nanoparticle. These components could impart sophisticated properties to the string if appropriately connected between electrodes and placed so that local fields could be varied. Such strings, or 3-dimensional networks, will be formed in-situ so as to have a fraction attached to the I/O electrodes.

Nanoparticles 100 are synthesized in different sizes, from ~1.5 nm to ~500 nm, as required. The surfaces will be activated, as seen in Fig. 9, by various chemical connecting molecules 108 to facilitate the interconnections. The final hybrid assemblies will be made by simply mixing activated components in non-aqueous solvents. The metal nanoparticles primarily are made of gold and these are synthesized at controlled sizes using established methods. Other similar conductive metals such as Pt, Pd, and Ag can also be used. Attachment of molecules such as RS-, RSe-, RTe-, and RNC- uses current attachment chemistry techniques, where R is connecting molecule 108 with variations chosen to optimize surface bonding and junction impedances. For Pt, the same groups can be used. The nanoparticles 100 are most easily made in an aqueous solution. However, for assembling the hybrid structures, non-aqueous conditions must be used in order to maximize molecular solubilities. This is done by pre-functionalizing the nanoparticle surface with short, solubilizing alkanethiolate chains (~C_4H_9-) or other coordination groups to form capped structures. These moieties will be transferred to appropriate organic solvents such as THF in which the molecular wires or switches will be attached by insertion into the host alkanethiolate surface matrix on the nanoparticle. This strategy will also serve to orient the inserted molecule in an outward direction from the surface. Electroactive molecules can be placed onto the surface of Au particles in the 1-2 nm size range and thus give those particles electrochemical effects.

Referring now to Fig. 10, using semiconductor nanoparticles such as CdS and CdSe or fullerenes (C_{60}) 112 and decorating their surfaces with the controller molecules 110, would permit the nanoparticles and fullerenes to act as switchable capacitors: charging in the presence of an applied field and retaining the charge once the field is diminished. C_{60} can be electrochemically reversibly charged with up to six electrons. The first four electrons are accepted with great facility (-0.98, -1.37, -1.87, and -2.37 volts vs. Fc/Fc+). Therefore, this could serve as an excellent molecular capacitor. The semiconductor particles consist primarily of metal chalcogenides, such as CdSe, CdTe, ZnSe, and ZnTe. The Cd-based nanoparticles are well-known and can be prepared with controlled sizes. Their synthesis and surface capping follows well-known established procedures. The surface functionalization of the chalcogenides involves primarily RS-, RSe-, and RTe- attachment groups. The electronic band structures of the capped nanoparticles and hybrid assemblies are probed using X-ray photoemission and optical spectroscopy. These particles can be attached to metal surfaces using self-assembly chemistry. For particles and hybrids assembled at surface, photoemission and spectroscopic ellipsometry are used. The physical structures can be examined by surface probes including STM and AFM and electron microscopes including TEM and high resolution SEM.

Insulating oxides, such as Al_2O_3, TiO_2, and SiO_2, are surface-capped with chemical groups for selective interconnection to conductive metal particles. These oxide particles act as high impedance resistors in the circuit. Control of the fractions of insulating particles in the self-assembling mixture will thus allow control of the total number of circuit connections up to the ultimate limit of the percolation threshold. Approaching this limit will impart exceedingly high sensitivity to the current-voltage characteristics across the individual I/O terminals in the device. The insulator nanoparticles are made using standard methods. Typically, TiO_2 and Al_2O_3 are used. The latter is desirable since hydrated Al_2O_3 surfaces can be very efficiently and selectively bonded and functionalized using organic acids such as RCO_2H and RPO_3H_2. The phosphonic acids are useful since they form a "tripod" bonding to the surface which orients the adjacent part of the adsorbate perpendicular to the surface (see Fig. 11). SiO_2 nanoparticles also make excellent insulator components but the best surface functionalization chemistry involves SiO_X bonding which forms polymeric networks and is difficult to control. Other possible nanoparticles include SnO_2. Functionalization chemistry is available, for example, RS- chemistry developed for indium tin oxide electrodes transfers to SnO_2.

Modular pre-assembly of the moleware, illustrated in Fig. 12, is done prior to the intelligent self-assembly within the container. All components are added as needed and the assembly permitted. If more memory is desired, more DRAMs are added, for example. By varying the chemical nature of the surfaces of each type of component nanoparticle, the relative interconnection affinities of the molecular and nanoparticle components can be controlled, thus providing a rational means to tune the final structural and functional states of the present modular computer. This strategy, together with the control of relative amounts of components, their addition sequence and intermittent real-time feedback-driven pulsing of the device I/O ports. forms the key to the overall intelligent self-assembly approach.

For example, there are a variety of strategies for assembling the functional hybrid strings and networks. The general approach is to cap each nanoparticle 120 by a layer of bifunctional molecules 122 (where "A" is -SH). In the cases of metals and semiconductors where molecular wires 124 will be used, wires 124 must be oriented away from the nanoparticle's surface by insertion into a pre-assembled host monolayer of a short chain molecule. For insulators, the bifunctional molecules 122 can be insulating so that the precapping step is not necessary. The attachment groups A, C, F for a gold nanoparticle 120 will typically be A = -SX (X = H, Ac, etc.) or –NC for Au, C = SX (X = H, Ac, etc.) for CdSe, etc. and F= be $-CO_2H$ or PO_3H for Al_2O_3. The surface termination groups will typically be B = A, D = C and G = A or C. A variety of starting nanoparticles 120 of different types, sizes and surface termination chemistry can be made. Varied relative amounts of desired types of the starting nanoparticles 120, including simple metal particles (e.g., Au with a "bare" surface), will be added, together and in sequences as desired, into a solvent containing a gold electrode surface functionalized with a molecular wire 124 terminated by B. Choosing surface terminations that attach to nanoparticle 120 will cause the mixture to interconnect in a statistical fashion both to itself in solution and to the Au electrode. These kinds of surface terminations will cause a network 136 to grow outward from the surface 140. A large number of possible paths will be produced which contain interconnections of metals 128, semiconductors 130 and insulators 132, thus generating a highly complex functional electrical network 136. Given the close proximity of neighboring base electrodes, network 136 will be established across the electrode gap forming a circuit connection, as illustrated in Fig. 12. Given the wide choices in the starting nanoparticle materials, their sizes, chemical terminations, the mixture compositions and mixing sequences, a huge variety of possibilities exists for controlling the statistical character of the network circuits. The next extensions of this strategy will include the addition of controller molecules to impart potential field-activated switching properties to the networks. The controller molecules can be added both by including them in the nanoparticle caps and by adding the molecules to the mixing solutions so that they become incorporated in the growing network. The structures of the networks assembled at single metal electrodes will be probed by elemental analysis (e.g., wet chemical techniques for the w, x, y, z stoichiometric factors), high resolution SEM, AFM and TEM (for cases of networks grown from surfaces such as SiO or C on TEM grids), and NMR.

Precisely defined molecular architectures including molecular wires 150, wires with tunnel barriers 152, wires with quantum wells 154, three terminal junctions 156, three terminal systems 158 that may act as switches or transistor, and four terminal molecules 160 with strategically-placed tunnel barriers that may act as logic devices are illustrated in Fig. 13. These are embodiments of the more general types of moleware illustrated in Fig. 1B, with molecular wire 30 corresponding to molecular wire 150, wire with tunnel barriers 32 corresponding to wire 152, wire with quantum well 34 corresponding to wire 154, three terminal junction 36 corresponding to junction 156, three terminal switch 38 corresponding to switch 158, and four terminal molecule corresponding to molecule 160. Note that "SAc" indicates a "protected aligator clip", which has an acetyl group that is removed during self-assembly permitting sulfur to form contacts with metal surfaces. The acetyl group is removed by adding a base such as ammonium hydroxide, although it can be removed by the gold surface itself, albeit more slowly.

All of the syntheses have been developed in a convergent fashion in solution or using synthetic methods on a polymer support to streamline the synthetic protocols. Using compounds similar to the wire with two tunnel barriers, we observed negative differential resistance as expected from a resonance tunneling diode. Therefore, precise molecular architecture has been used to build device-like properties into nanoscale systems. Molecules such as those shown can be used in the present molecular computer array.

Described below is a moleware preparation, deposition on a two-dimensional chip platform, and subsequent test of the I/O leads.

Moleware preparation: A solution of octadecyl-terminated single-walled carbon nanotubes (1) was prepared by a slight modification of a procedure by Chen, J. et al. (Chen, J.; Hamon, M. A.; Hu, H.; Chen, Y.; Rao, A.; Ecklund, P. C.; Haddon, R. C. Science 1998, 282, 95-98). To 100 mg of crude single-walled carbon nanotubes (SWNTs, obtained from Tubes@Rice) was added concentrated sulfuric acid/nitric acid (3:1 v/v, 10 mL). The mixture was sonicated for 4 h (Cole-Parmer, model B3-R) followed by filtration (0.2 micron PTFE) and washing with methanol. To the resulting oxidized SWNTs (14.5 mg) was added thionyl chloride (25 mL, 343 mmol) and dimethylformamide (1 mL). The mixture was sonicated for 0.5 h and heated to 70 C for 24 h. After cooling and centrifugation (Fisher Scientific, Marathon 22 K. 4200 PRM) for 2 min, the solid was collected by filtration and washed with freshly distilled THF and then dried under vacuum (6 mm of Hg) at room temperature. A mixture of the resulting SWNTs (13 mg) and octadecylamine (2 g, 7.4 mmol) was heated to 100 C for 4 days. After cooling to room temperature, the solid was washed with ethanol, sonicated with diethyl ether/hexane (1:1, v/v), collected by centrifugation and filtration and dried under vacuum (6 mm of Hg) at room temperature for 1 h. Some exogenous octadecylamine remained with 1. The resulting 1 could be dissolve in chloroform, tetrahydrofuran, or methylene chloride.

Moleware Assembly on Chip: An array of 60 gold dot electrodes, as illustrated in Fig. 14, (10 micron diameter per dot, 100 micron pitch distance between dots) had one fourth of the dots exposed (15 dots), the remainder of the dots were coated with titanium dioxide, and hence inactive. Each of the electrode dots was affixed to a bonding pad via a titanium dioxide-coated gold wire. A solution of 1 in chloroform (0.01 w/w) was prepared. A 200-mg portion of the solution of 1 in chloroform was placed over the entire chip. After air drying, the chip was placed under vacuum (6 mm of Hg) at room temperature for 15 min. The deposition was somewhat ragged and the film cracked.

Electrical Measurements on Chip.

The resistance between any two adjacent electrodes was generally below 20 Kohms, typically 12 Kohms, as measured by a multimeter. One electrodes pair had a resistance of ~100 Kohms between them.

Zero-bias conductivities of ~10-100 micro Siemens were recorded.

I(V) traces (current-voltage) show nonlinearity around zero, and high field saturation. The low bias nonlinearity may be due to the 1-1 junctions, although further characterization would be necessary. The high field noise is higher than at the low field.

Using third terminal current injection, the conductivity (between two electrodes) could be modified (remotely, by the third electrode) downward in most cases. This is rather reminiscent of degradation and breakdown. However, occasionally, an increase in conductivity (~2 fold) was noted, with some hysteretic behavior, reminiscent of electrical gain.

It will be apparent to those skilled in the art of nanotechnology and computer architecture that many substitutions and modifications can be made to the preferred embodiments described above without departing from the spirit and scope of the present invention, defined by the following claims.

WHAT IS CLAIMED IS:

1. A molecular computer, comprising:

a container formed of a lattice having

a multiplicity of input pins carried by said container,

a multiplicity of output pins carried by said container; and

moleware contained within said container.

2. The molecular computer as recited in claim 1, wherein said moleware is selected from the group consisting of metal particles for interconnecting said moleware, molecular alligator clip-bearing 2-, 3-, and 4-terminal molecular wires, carbon nanotubes, molecular resonant tunneling diodes, molecular switches, molecular controllers, molecular DRAM and SRAM components selected from the group consisting of molecular controller/nanoparticles and fullerene hybrids, and combinations thereof.

3. A molecular computer made by a process comprising the steps of:

attaching plural input and output pins to a containment; and

injecting moleware into said containment;

allowing said moleware to form bridges between said input and said output pins;

applying voltages to input pins;

measuring voltages at said output pins; and

identifying the relationships between input and output pins when voltages are applied to said input pins until a truth table is constituted for said computer.

4. The molecular computer as recited in claim 3, wherein said moleware is selected from the group consisting of metal particles for interconnecting said moleware, molecular alligator clip-bearing 2-, 3-, and 4-terminal molecular wires, molecular resonant tunneling diodes, carbon nanotubes, molecular switches, molecular controllers, molecular DRAM and SRAM components composed of molecular controllers or fullerene hybrids, and combinations thereof.

5. The molecular computer as recited in claim 3, wherein said applying step is done by applying said voltage to a bundle of said input pins and said measuring step is done by measuring voltages at a bundle of said output pins.

6. The molecular computer as recited in claim 3, further comprising the step of applying field gradients across said containment to modify said bridges formed by said moleware between said input and said output pins.

7. The molecular computer as recited in claim 3, further comprising the step of electrochemically inducing cross linking of moleware to modify bridges formed by said moleware between said input and said output pins.

8. The molecular computer as recited in claim 3, further comprising the step of eliminating a portion of said bridges formed by said moleware between said input and said output pins.

9. The molecular computer as recited in claim 3, wherein said moleware components are less than 1000 nanometers in diameter.

10. The molecular computer as recited in claim 4, wherein said injecting step is done by injecting each type of moleware separately and at intervals.

11. The molecular computer as recited in claim 3, further comprising the step of installing a support matrix inside said containment prior to injecting said moleware.

12. The molecular computer as recited in claim 3, further comprising the step of varying said voltages applied to said input pins until a desired output at said output pins is achieved.

13. The molecular computer as recited in claim 3, wherein a portion of said moleware includes semiconductor particles coated with molecular control elements adapted to perform memory functions for said molecular computer.

14. The molecular computer as recited in claim 3, wherein said identifying step further comprises the steps of:

connecting a computer to said input pins and said output pins; and

using said computer for identifying said relationships.

15. The molecular computer as recited in claim 3, further comprising the step of programming said molecular computer using error resilient software.

16. The molecular computer as recited in claim 3, further comprising the step of synthesizing said moleware in solution before injecting said moleware.

17. The molecular computer as recited in claim 4, wherein said molecular wires are constructed from materials selected from the group of oligo(phenylene ethynylene)s.

18. The molecular computer as recited in claim 17, wherein said molecular wires have termini made of a material selected from thiol, planar oligomer segments or carbon nanotubes or bundles thereof.

19. The molecular computer as recited in claim 17, wherein said molecular wires are selected from the group consisting of carbon nanotubes, planar polyphenylenes, polypyridines, polypyrazines and polythiophenes.

20. The molecular computer as recited in claim 4, wherein said particles are made of a material selected from gold, silver, palladium, platinum and alloys thereof.

21. The molecular computer as recited in claim 3, wherein said solution is a nonaqueous solution.

22. The molecular computer as recited in claim 20, wherein said particles are capped with bifunctional molecular wires.

23. The molecular computer as recited in claim 4, wherein said molecular switches are selected from the group consisting of semiconductor particles having controller molecules on their surfaces and fullerenes.

24. The molecular computer as recited in claim 23, wherein said semiconductor particles are selected from the group consisting of CdS and CdSe.

25. The molecular computer as recited in claim 4, wherein said molecular resistors are made of insulating oxides adapted for connection to metal particles.

26. The molecular computer as recited in claim 25, wherein said insulating oxides are selected from the group consisting of Al_2O_3, TiO_2, and SiO_2.

27. The molecular computer as recited in claim 4, wherein said alligator clips are made of a transition metal attached to a metallic surface.

28. The molecular computer as recited in claim 4, wherein said molecular controllers each have a zwitterionic form that stabilizes in the presence of an applied electromagnetic field.

29. The molecular computer as recited in claim 28, wherein said molecular computer is biphenylnitroamine.

30. The molecular computer as recited in claim 3, wherein said solution is an organic solvent.

31. The method as recited in claim 3, further comprising the step of adapting said moleware to increase interconnection affinities of said moleware to each other.

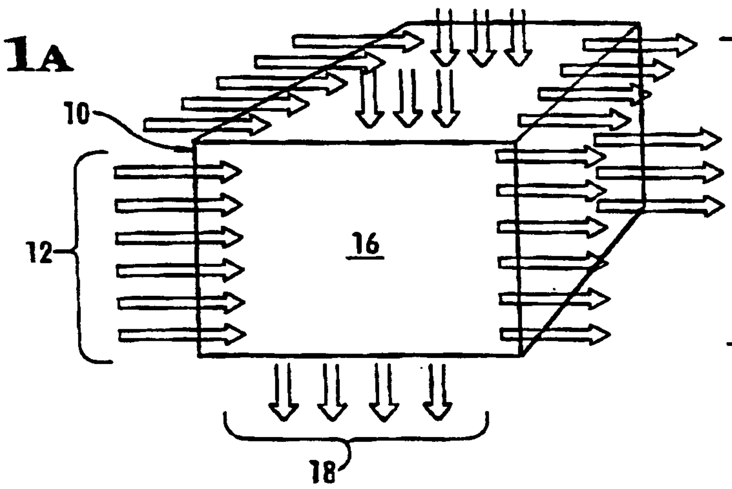
32. The method as recited in claim 31, wherein adapting step is selected from the group consisting of capping, functionalizing, activating and combinations thereof.

33. A method of transferring data from an input to an output, said method comprising: establishing a molecular bridge composed of molecules between an input and an output, said molecules including a first molecule connected to said input; and applying a voltage to said input sufficient to modify the molecular electrostatic potential of a first molecule.

34. The method as recited in claim 33, further comprising the step of adapting said molecules to increase interconnection affinities of said molecules to each other.

35. The method as recited in claim 34, wherein adapting step is selected from the group consisting of capping, functionalizing, activating and combinations thereof.

FIG. 1A



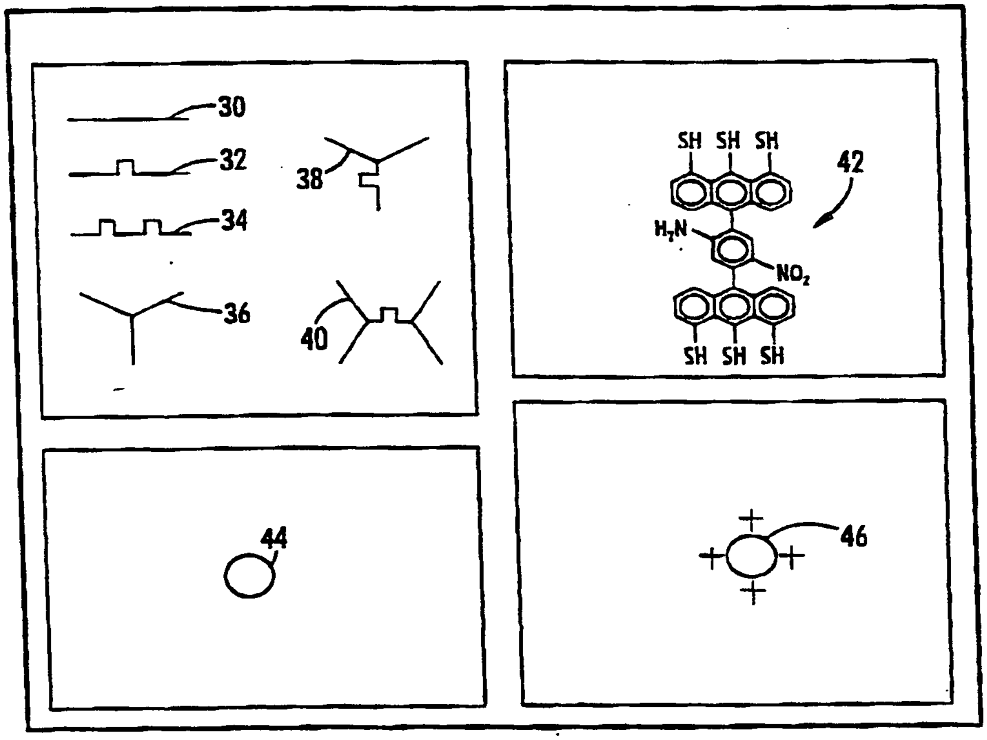


FIG. 1B

FIG. 2A

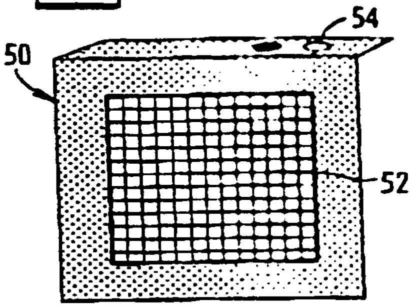


FIG. 2B

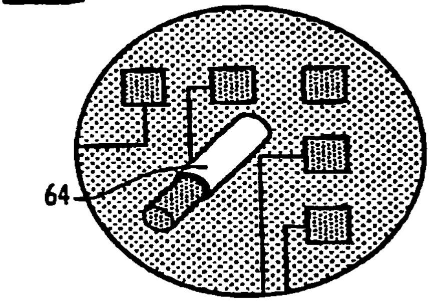
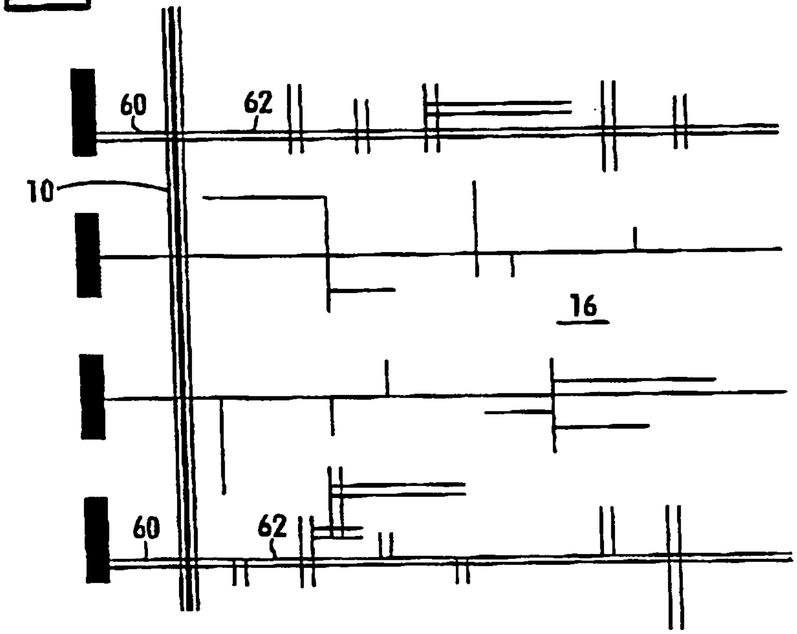


FIG. 3



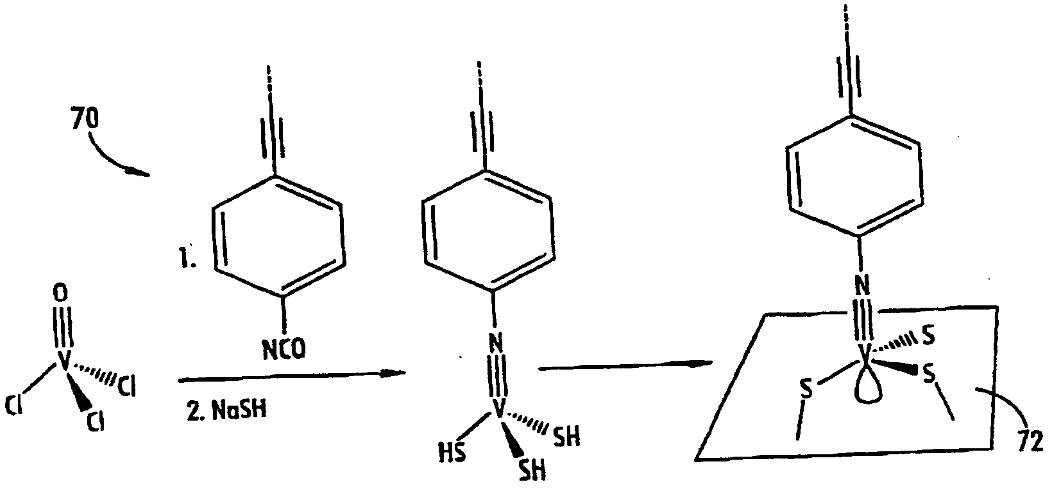
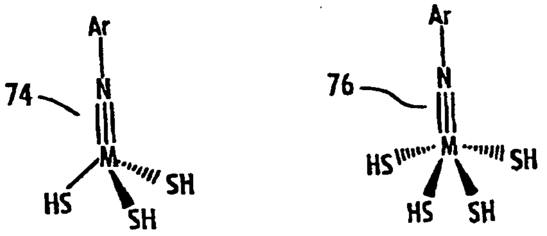

70
1.
NCO
O
V
Cl
Cl
Cl
2. NaSH
N
V
HS
SH
SH
N
V
S
S
S
72

Fig. 4

Fig. 5


Ar
N
74
M
HS
SH
SH
Ar
N
76
M
HS
SH
HS
SH

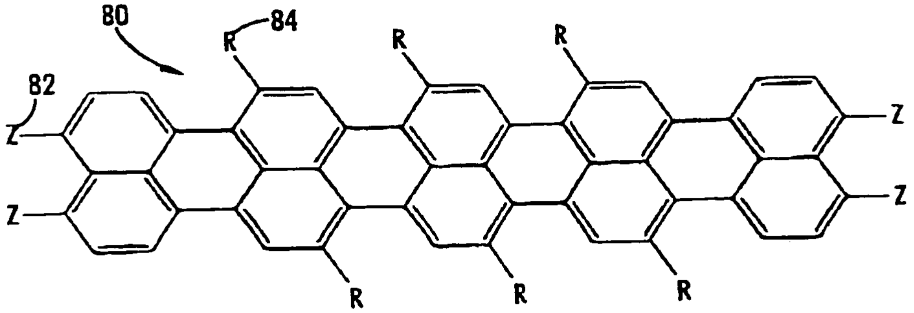


FIG. 6

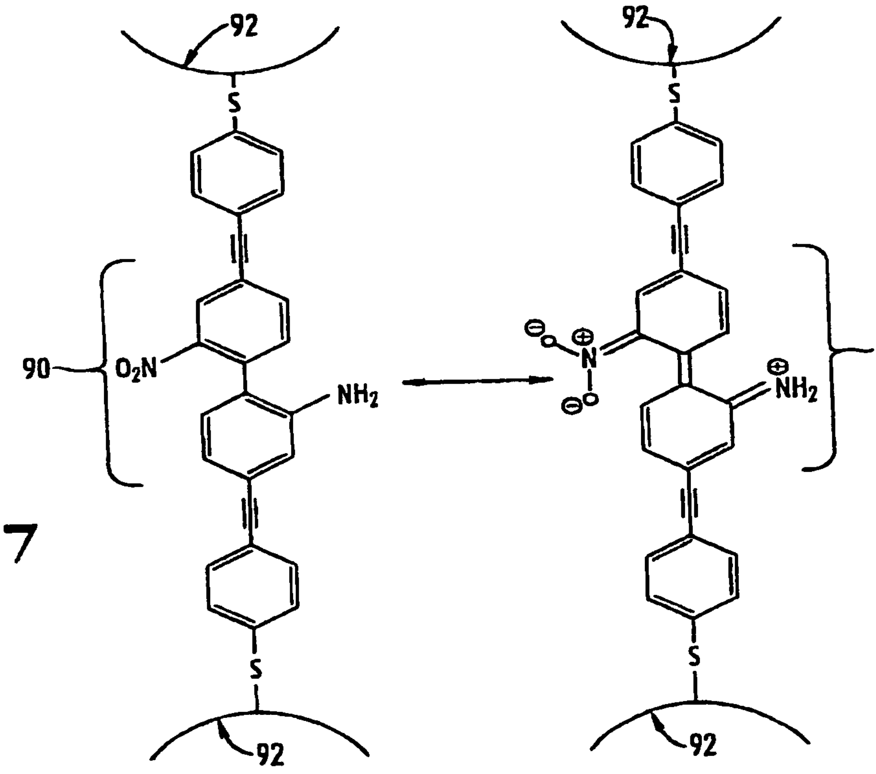


FIG. 7

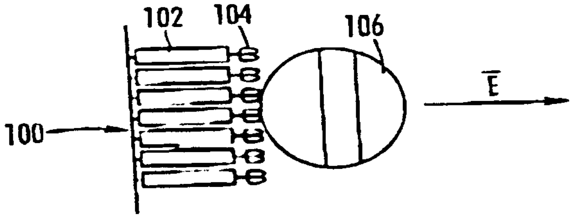


FIG.8

FIG.9

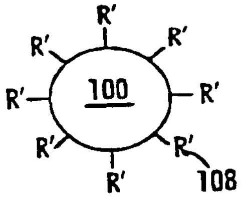


FIG.10

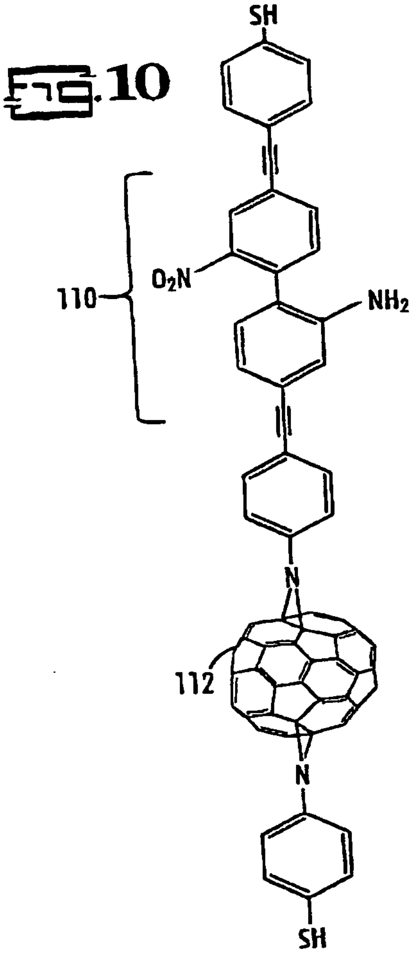
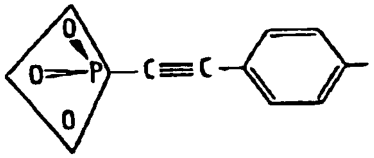


FIG.11



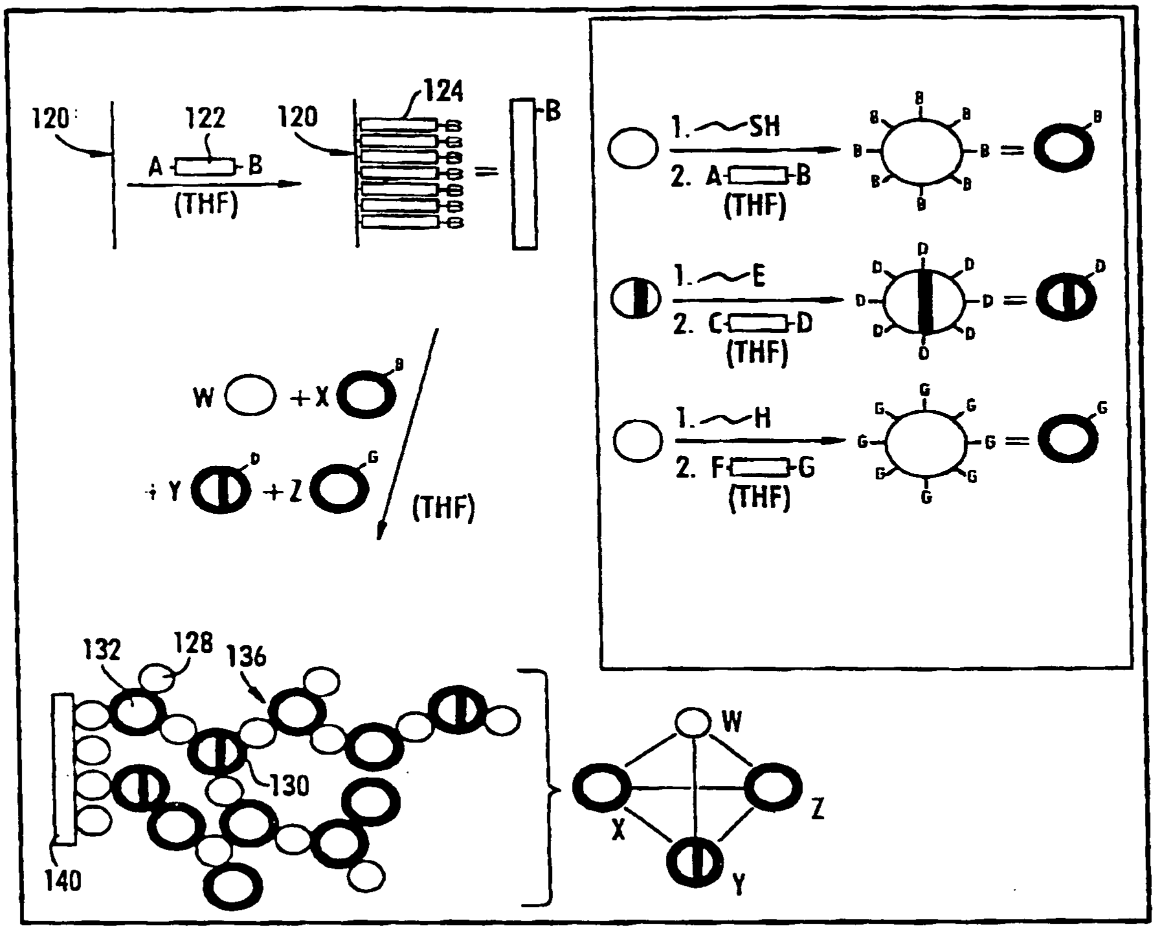
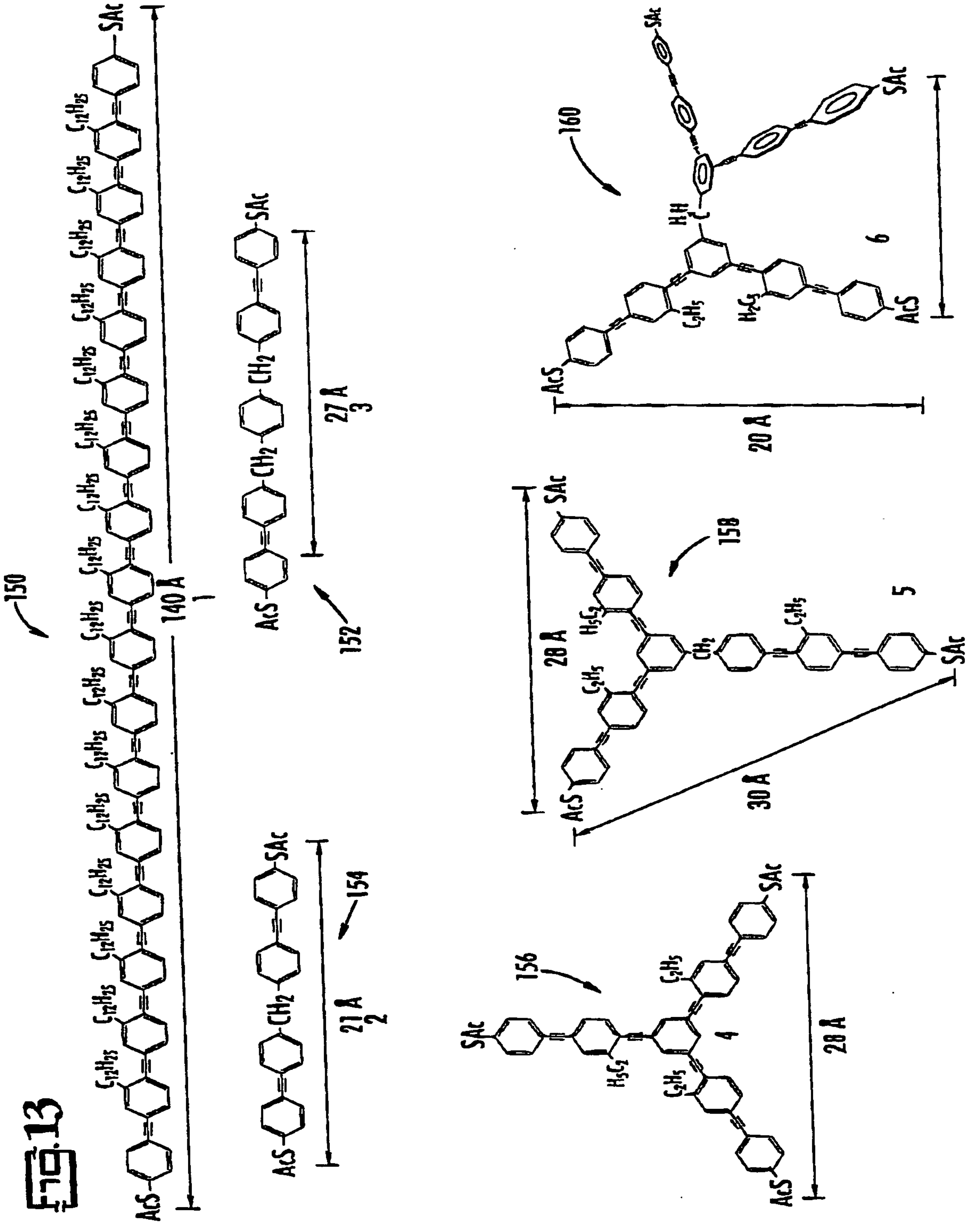

120
122
A
B
(THF)
120
124
B
1. SH
2. A B
(THF)
B
1. E
2. C D
(THF)
D
1. H
2. F G
(THF)
G
W
+X
B
+Y
D
+Z
G
(THF)
132
128
136
130
140
W
X
Y
Z

FIG. 12

Fig. 13


150
AcS
$C_{12}H_{25}$
SAc
140 Å
1
152
AcS
CH_2
SAc
27 Å
3
154
AcS
CH_2
SAc
21 Å
2
156
SAc
H_5C_2
C_2H_5
C_2H_5
AcS
SAc
4
28 Å
158
SAc
H_5C_2
C_2H_5
CH_2
C_2H_5
AcS
SAc
28 Å
30 Å
5
160
SAc
H H
C
C_2H_5
H_5C_2
AcS
AcS
SAc
6
20 Å

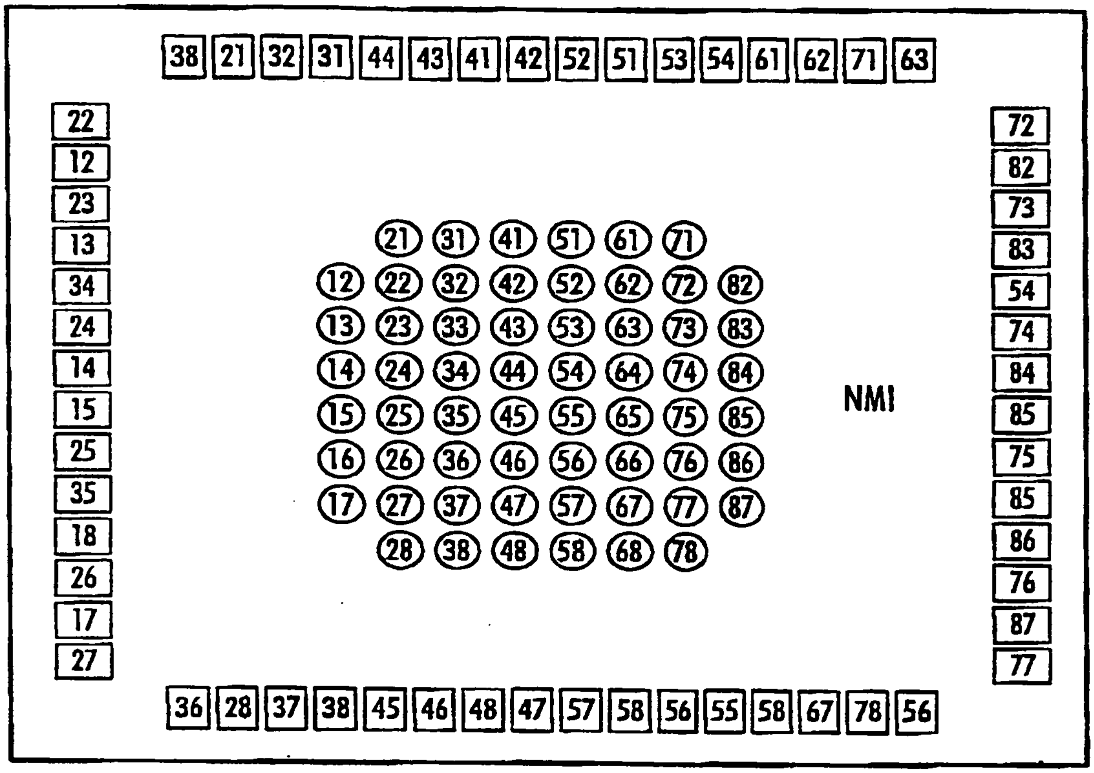

38 21 32 31 44 43 41 42 52 51 53 54 61 62 71 63
22
12
23
13
34
24
14
15
25
35
18
26
17
27
21 31 41 51 61 71
12 22 32 42 52 62 72 82
13 23 33 43 53 63 73 83
14 24 34 44 54 64 74 84
15 25 35 45 55 65 75 85
16 26 36 46 56 66 76 86
17 27 37 47 57 67 77 87
28 38 48 58 68 78
NMI
72
82
73
83
54
74
84
85
75
85
86
76
87
77
36 28 37 38 45 46 48 47 57 58 56 55 58 67 78 56

Fig.14

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US00/01360

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) : H03K 19/00, G06F 9/455, 17/50; G06G 7/48
US CL : 703/4, 11, 13, 23, 716/17
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S 703/4, 11, 13, 23, 716/17

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Science Server IEEE EAST

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5,401,975 A (IHARA et al.) 28 March 1995, fig. 24-30, col. 10-13	1-35
Y	US 5,589,692 A (REED) 31 December 1996, fig. 4, 6, col. 5-7, 15-38	1-35
Y	Draganescu, M., "From Solid State to Quantum and Molecular Electronics, the Deeping of Information Processing"; October 1997; 1997 Int. Semiconductor Conf.; abstract, pp. 16-18.	1-35
X	Conrad, M.; "Molecular Computing: the Lock-key Paradigm"; Computer; November 1992 .Vol. 25. No. 11. pp. 12-20	1-35

□ Further documents are listed in the continuation of Box C. □ See patent family annex.

*	Special categories of cited documents	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date		
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed	"&"	document member of the same patent family

Date of the actual completion of the international search 21 MAY 2000	Date of mailing of the international search report 12 JUN 2000
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer KEVIN TESKA James R. Matthews Telephone No. (703) 305-9704

US005706404A

United States Patent [19]

Colak

[11] **Patent Number: 5,706,404**

[45] **Date of Patent: Jan. 6, 1998**

[54] **NEURAL NETWORK USING INHOMOGENEITIES IN A MEDIUM AS NEURONS AND TRANSMITTING INPUT SIGNALS IN AN UNCHANNELLED WAVE PATTERN THROUGH THE MEDIUM**

[75] Inventor: **Sel B. Colak**, Eindhoven, Netherlands

[73] Assignee: **U.S. Philips Corporation**, New York, N.Y.

[21] Appl. No.: **688,462**

[22] Filed: **Jul. 30, 1996**

Related U.S. Application Data

[62] Division of Ser. No. 201,609, Feb. 25, 1994.

[30] **Foreign Application Priority Data**

Mar. 3, 1993 [EP] European Pat. Off. 93200603

[51] **Int. Cl.^{6}** **G06E 1/00**; G06F 15/18

[52] **U.S. Cl.** **395/24**; 395/23; 395/21

[58] **Field of Search** 595/20–25, 27; 382/155–159

[56] **References Cited**

U.S. PATENT DOCUMENTS

5,165,010	11/1992	Masuda et al.	395/27
5,426,720	6/1995	Bozich et al.	395/22
5,428,711	6/1995	Akiyama et al.	395/25

OTHER PUBLICATIONS

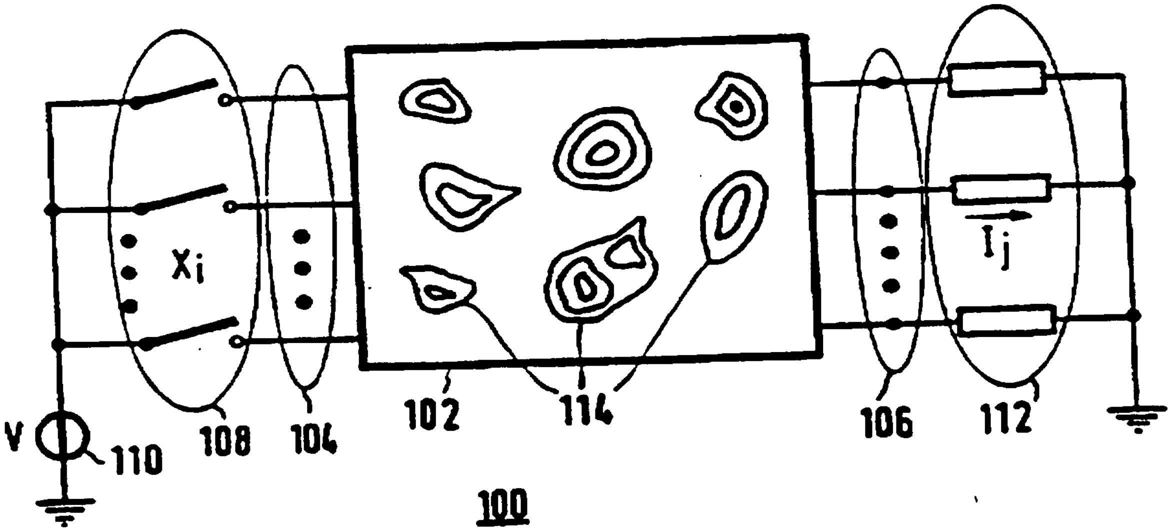
Ohta et al. "Variable sensitivity photodetector for optical neural networks"; Journal of Lightwave Technology, vol. 9, iss. 12, pp. 1747–1754, Dec. 1991.

Primary Examiner—Tariq R. Hafiz
Attorney, Agent, or Firm—Anne E. Barschall

[57] **ABSTRACT**

Neural net with spatially distributed functionalities. An information processing system comprises a neural net with fully distributed neuron and synapse functionalities in a spatially inhomogeneous medium to propagate a response field from an input to an output. The response field is a reaction of the medium to a plurality of input signals and depends non-linearly on the input signals. The response field is also determined by the inhomogeneities. The value of the field at one or more particular locations is indicative of one or more output signals of the neural net.

9 Claims, 5 Drawing Sheets



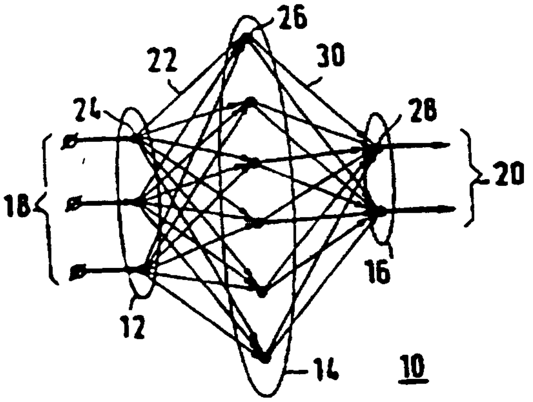


FIG.1
PRIOR ART

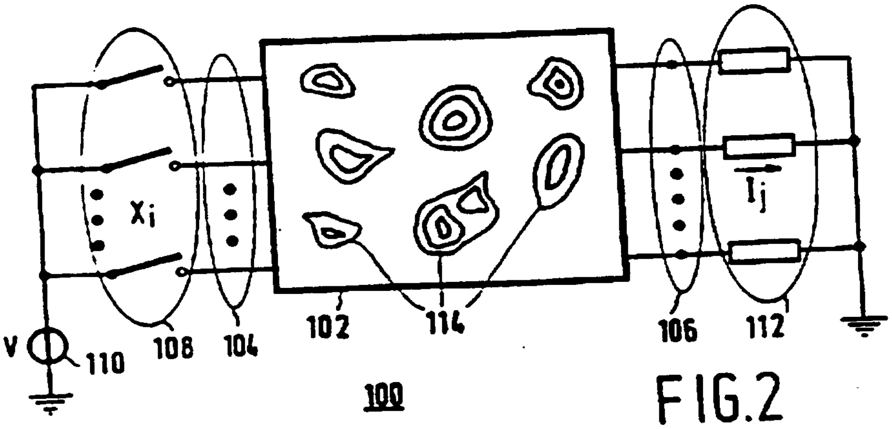


FIG.2

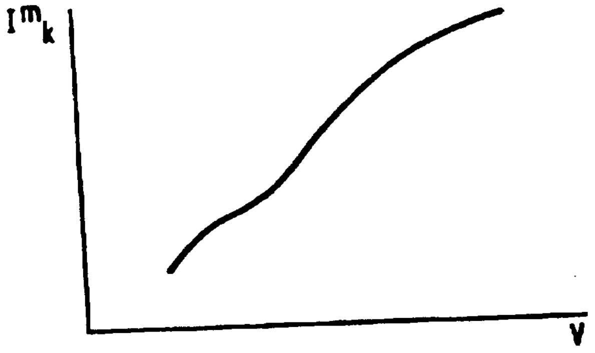


FIG.3

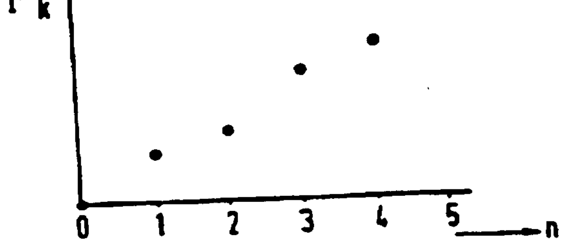


FIG.4

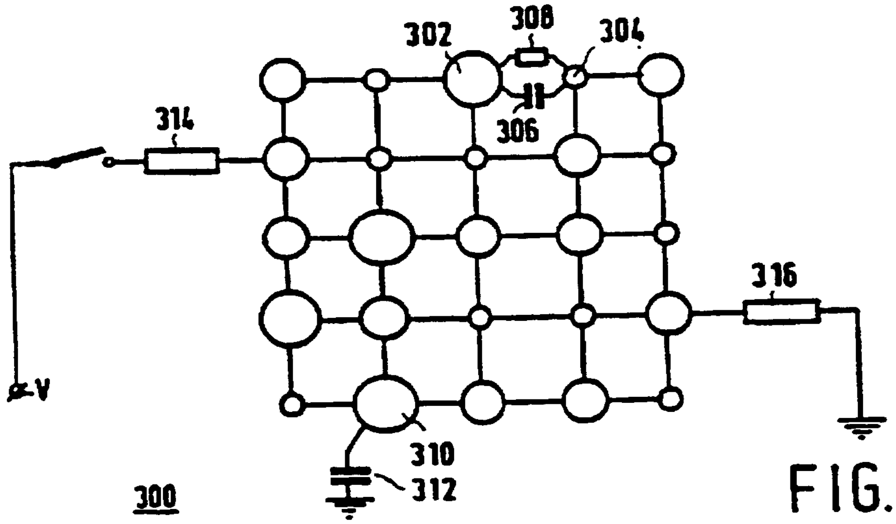


FIG.5

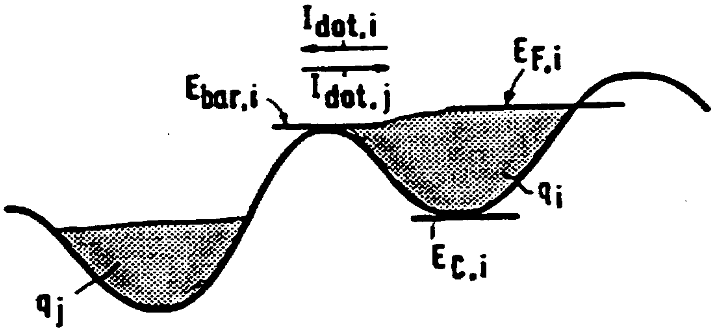


FIG.6

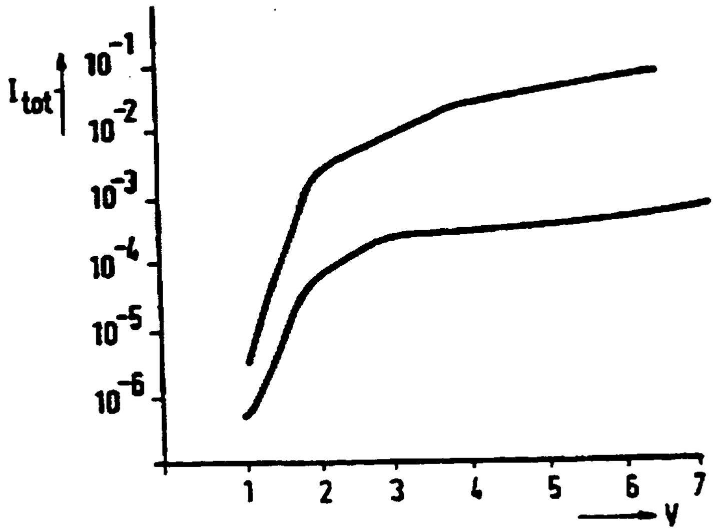


FIG.7

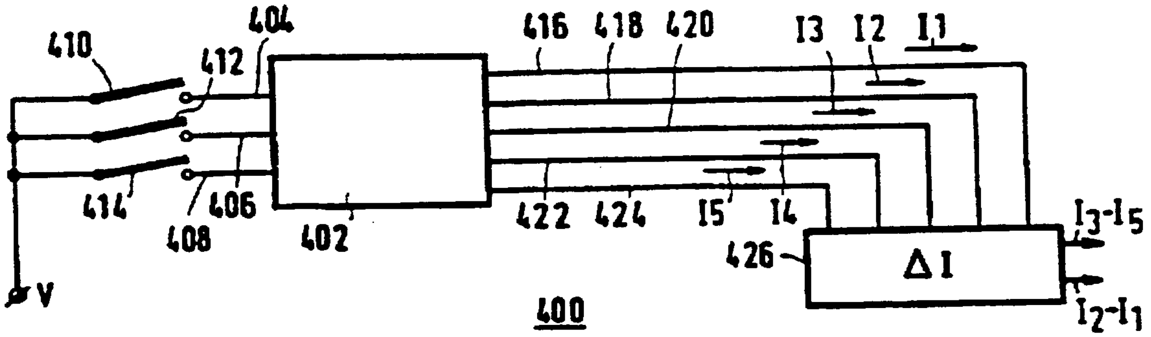


FIG.8

X	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}0\\1\\0\end{pmatrix}$	$\begin{pmatrix}1\\0\\0\end{pmatrix}$	$\begin{pmatrix}0\\1\\1\end{pmatrix}$	$\begin{pmatrix}1\\0\\1\end{pmatrix}$	$\begin{pmatrix}1\\1\\0\end{pmatrix}$	$\begin{pmatrix}1\\1\\1\end{pmatrix}$
I_1	0	3,8	3,9	6,8	6,9	7,8	10,5
I_2	0	8,1	8,4	10,1	10,3	10,6	12,2
I_3	0	0,8	0,9	3,5	3,6	4,6	7,5
I_4	-	-	-	-	-	-	-
I_5	0	0,9	1,1	2,8	2,9	3,7	6,0

Table I

FIG.9

X	$\begin{pmatrix}0\\1\\0\end{pmatrix}$	$\begin{pmatrix}1\\0\\0\end{pmatrix}$	$\begin{pmatrix}0\\1\\1\end{pmatrix}$	$\begin{pmatrix}1\\0\\1\end{pmatrix}$	$\begin{pmatrix}1\\1\\0\end{pmatrix}$	$\begin{pmatrix}1\\1\\1\end{pmatrix}$
I_2-I_1	4,3	4,5	3,3	3,4	2,8	1,7
I_5-I_3	0,1	0,2	-0,7	-0,7	-0,9	-1,5

Table II

FIG.10

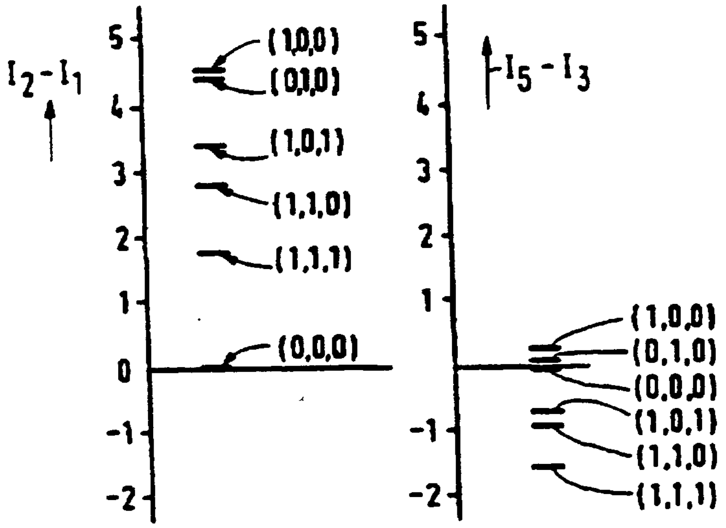


FIG.11

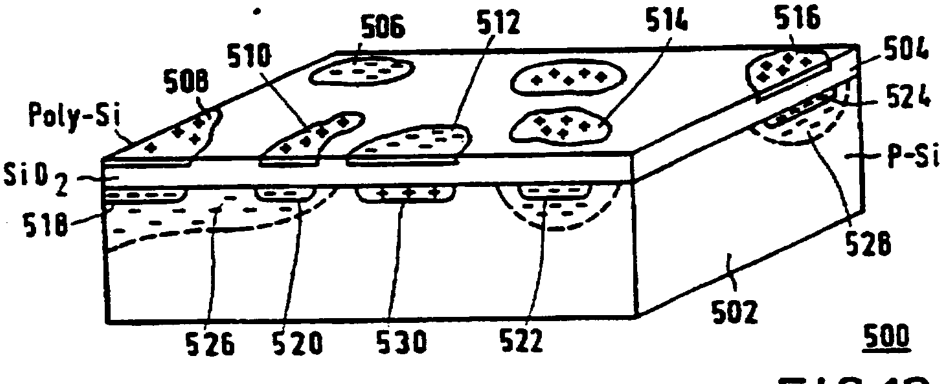


FIG.12

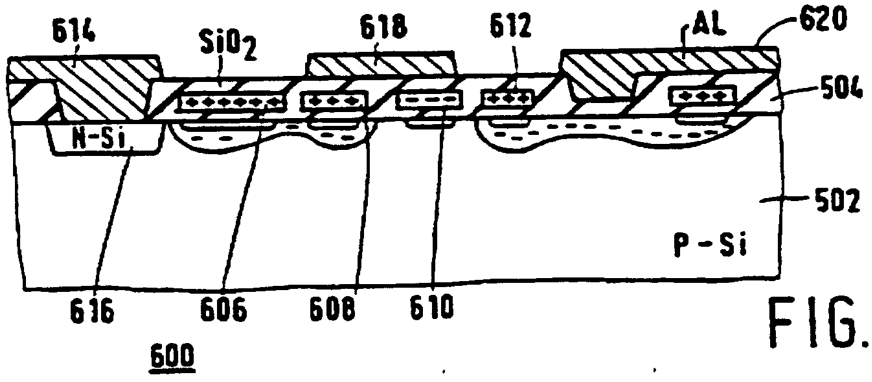


FIG.13

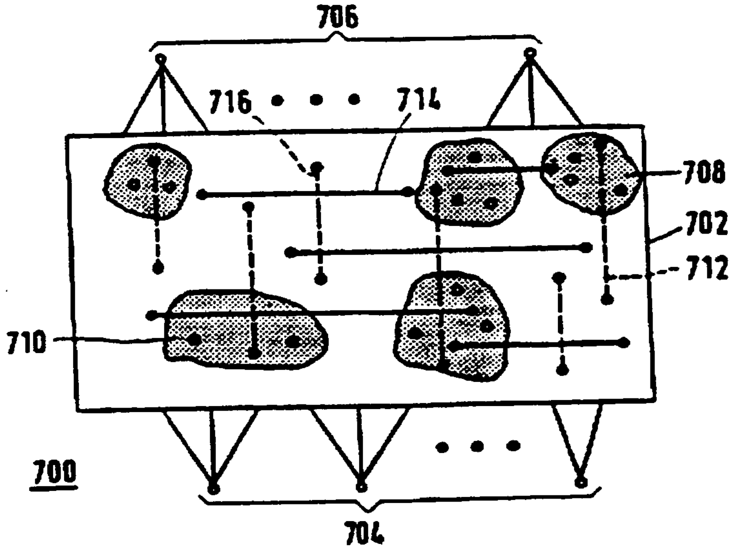
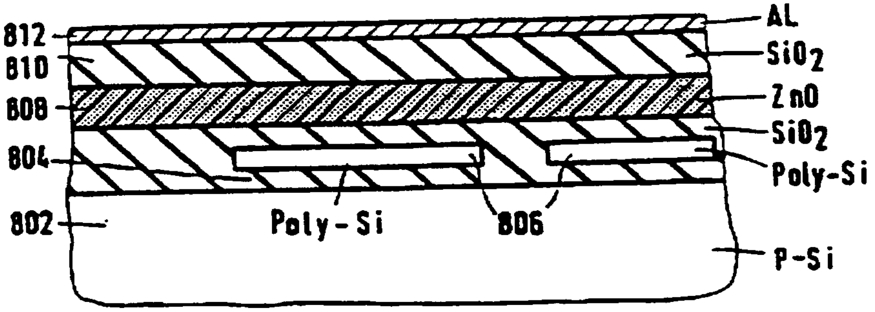


FIG.14



800

FIG.15

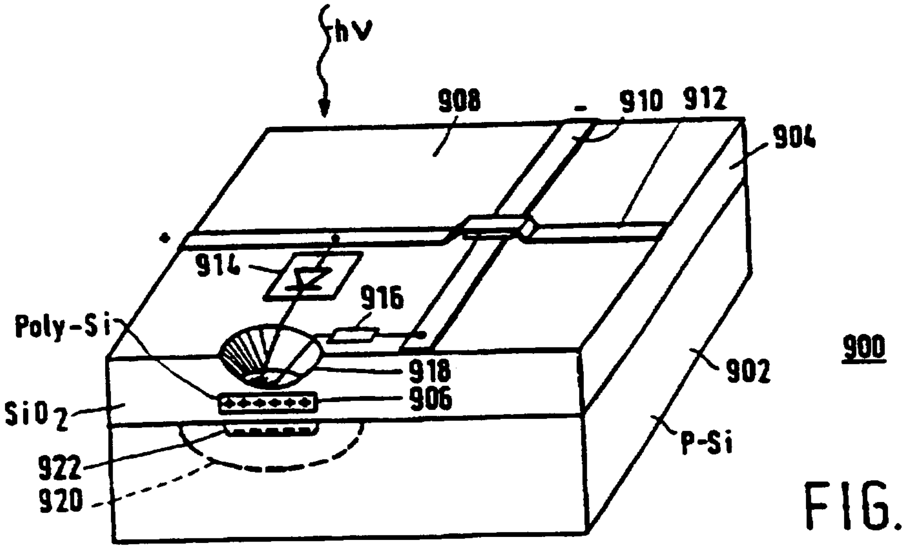


FIG.16

NEURAL NETWORK USING INHOMOGENEITIES IN A MEDIUM AS NEURONS AND TRANSMITTING INPUT SIGNALS IN AN UNCHANNELLED WAVE PATTERN THROUGH THE MEDIUM

This is a division of application Ser. No. 08/201.609. filed on Feb. 25, 1994.

FIELD OF THE INVENTION

The invention relates to an information processing system with a neural net functionally composed of neurons interconnected by synapses. The network has an input means to receive a plurality of input signals, an output means to provide at least one output signal, and an arrangement between the input means and the output means to implement a neural transformation of the plurality of input signals for generating the output signal.

BACKGROUND ART

Over the last decade, neural networks have attracted an ever increasing interest as signal processors. Such processors essential characteristics reside in the non-linear mapping of input signals onto output signals using spatially distributed elements and massively parallel information processing in a densely interconnected structure. Such a processor is a robust and fault-tolerant processing device. Furthermore, neural networks can be programmed by means of training on the basis of examples rather than by means of prescribing algorithmic instructions such as in the field of conventional microprocessors. Training can be accomplished with or without a supervisor, the latter version being called "self-learning". Neural networks are expected to play an important role, especially in the fields of associative memories, classifiers, pattern recognition and optimization problems. For a general introduction, see: "An Introduction to Computing with Neural Nets", R. P. Lippmann, IEEE ASSP Magazine, April 1987, pp. 4–22.

Although some of the mathematical algorithms created for neural network control or neural network simulation appear to be useful in seeking software solutions to particular problems, the real benefit of the neural network approach is believed to be obtainable only by way of dedicated hardware. A variety of electronic and optical hardware implementations of neural networks have seen the light over the last decade. For example, the Laboratoires d' Électronique de Philips has developed the first commercially electronic, fully digitally operating, trainable neural chip for modular neural network architectures, see U.S. Pat. No. 4,994,982. An electronic implementation of an analog neural net is known from, for instance, U.S. Pat. No. 4,866,645. Optical neural nets are dealt with in: "Optical Neural Computers", Y. S. Abu-Mostafa and D. Psaltis, Scientific American 256, March 1987, pp. 66–73.

Often, spatially tight designs are required to realize a sufficiently large number of the neurons and synapses in a confined space. A trainable neural net requires additional circuitry to individually and accurately modify the weights in the net's learning phase, necessarily leading to an increase of the system's size and complexity. Since the processing capability of a neural net increases with larger numbers of neurons and interconnections, nowadays efforts in neural net technology find a major target in increasing the density of neurons and interconnections. As an example, a major problem encountered in the design of electronic embodiments is the limitation to the number of neurons or neuron functionalities that can be implemented on a single chip. Typically, the maximum number of neurons or neuron functionalities that can be integrated into a chip using state of the art technology is well lower than one thousand. Furthermore, the complexity and the mount of additional components needed to implement the learning capability of a net containing such large numbers of neurons is expected to be at least proportionally higher.

The examples mentioned above have in common that the associated architectures primarily have a lumped character. That is, the various respective functional elements of the neural net which perform the basic computations, such as the neurons and the synapses, are implemented by physically distinct devices or circuits. These devices and circuits are dedicated to operations such as weight multiplication, summation and applying a sigmoid function.

One of the ways to reduce of the number of neurons or neuron layers, to simplify the interconnection scheme and to increase efficiency, is to employ high-order terms (squares, products, cubes, etc.) of the input signals supplied to the neural net. These high-order terms then serve as the signals to be handled by the neurons. See, for instance, C. Lee Giles et at., "Learning, Invariance and Generalization in High-Order Neural Nets", Appl. Opt., Vol. 26, 1987, pp. 4972–4978. The high-order dependencies are determined in advance and can be thought of as pre-processed input data, the preprocessing being based on some prior knowledge of the kind of problems that the neural net is meant to solve. For example, different classification problems to be handled by a neural net may require different forms of non-linear correlations between the input signals.

OBJECT OF THE INVENTION

In view of the limitations discussed above, it is therefore an object of the invention to provide an information processing system comprising a radically new neural net architecture, which permits integration of larger numbers of neuron functionalities and is highly technology-independent.

SUMMARY OF THE INVENTION

To this end, the invention provides an information processing system having a neural net that is functionally comprised of neurons interconnected by synapses. The net has an input means to receive a plurality of input signals, an output means to provide at least one output signal, and an arrangement between the input means and the output means to implement a neural transformation of the plurality of input signals for generating the output signal. According to the invention, the arrangement comprises a physical medium operative to propagate a response field throughout the medium in response to the plurality of input signals, the response field being non-linearly dependent on at least one of the input signals. The medium has one or more spatially distributed inhomogeneities operative to affect a spatial dependence of the field. The medium is operative to produce the output signal upon coupling at least a first response, representative of the field occurring at a first location in the medium, to the output means.

The invention is based on the insight that the energy level of the medium's ground state depends on the boundary conditions applied. Upon exciting the medium under respective different boundary conditions will lead to the medium settling in respective different lowest states. In this way, the boundary conditions serve as an input signal and the associated lowest energy configuration serves as an output signal, whose non-linear relation to the input signal is determined by the specific boundary conditions applied. In this manner, the mapping required for neural net operation can be attained using fully and substantially continuously distributed and collective physical properties of a spatially inhomogeneous medium. The medium may be an amorphous or a structured material, and may include a solid, a liquid, a gas, a plasma, etc. The relevant properties are associated with, for example, electrical, magnetic, electrodynamic or acoustic fields, or with phenomena involving the conversion from, e.g., an electric field into an acoustic field, etc.

The non-linear mapping and the mutual signal correlation are achieved in a spatially fully distributed and parallel fashion without necessitating clearly defined locations for neuron functionalities and synapse functionalities. This enhances the characteristic robustness proper to neural nets even further and, more importantly, avoids the spatial limitations inherent in conventional lumped systems. Consequently, higher densities of neuron functionalities and synapse functionalities are attainable in the invention. In addition, implementing the invention relies on the possibility of tailoring the non-linearities required in neural net operation to fit the properties achievable from various materials and devices. Since all materials exhibit in some way or another a nonlinear behavior, the invention considerably extends the field of suitable technologies to functionally realize a neural net.

Essential aspects of the neural net's internal operation include signal inhibition and signal crossing. Inhibition is a signal operation to decrease or reverse the signal's contribution to the collective behavior. Signal crossing occurs when two or more input signals each individually affect two or more output signals. In case the nonlinear characteristic is substantially monotonic within the parameter range used, additional measure may be required to obtain inhibition and signal crossing.

In case the medium's stimulus-response characteristic is substantially monotonic, the arrangement comprises encoding means coupling the physical medium to the output means for producing the output signal upon mutually correlating the first response and a second response. The operation of the encoding may involve, for example, a difference between a pair of responses, a ratio of a respective response and an aggregate sum of responses, or a change of a respective response brought about by a change of a particular one of the input signals.

Input signals may have different types. For example, one input signal may be of a dc voltage or of the energy of a photon, another one may be a current or an ac voltage.

The arrangement may be provided with a connection means between at least a second location and a third location in the medium to enable an interaction between the response field at the second location and the response field at the third location. Such an interconnection means permits to circumventing short range effects due to screening. Screening effects will occur in systems that are large in relation to the range of certain interactions. These effects will adversely affect the discrimination between responses to different sets of input signals. Therefore, beyond a given size of the system in the invention, resolution and noise become important. One way of compensating for these drawbacks is to use more contacts to extract more responses that combined produce the output signal(s). Another way is to provide a few long-range interconnections.

Preferably, the arrangement comprises a control means to selectively modify the spatially distributed inhomogeneities

with regard to at least position, size or strength. As mentioned the inhomogeneities interact with the continuous field in the medium. Selectively modifying the inhomogeneities provides a learning feature to the neural net of the invention.

An implementation of a neural net in the invention is a two-dimensional electronic transport layer composed of distributed, non-linear conduction paths between input contacts and output contacts. The transport layer can be implemented by use of sub-threshold conduction of semiconductor field-effect devices or by use of conducting granular layers of organic material or superconducting material. In a semiconductor embodiment, the inhomogeneities may be formed by inversion layers and depletion layers brought about by charged floating gates. By means of a mechanism that modifies the charge trapped at such a floating gate, the inhomogeneities can be adapted.

Another embodiment utilizes a non-linear optical system instead of the non-linear electronic transport layer discussed in the previous paragraph. Such a system then may be based on, for example, generating speckle patterns, i.e., interference patterns brought about by scattering, or second harmonic patterns in a material with an light-intensity dependent index of refraction when subjected to laser light.

It is noted that prior art literature associates a Hopfield neural network with spin glass systems. Spin glass systems assume macroscopic states based on the collective behavior of microscopic spins. Once excited, the spin glass traverses an extremely long sequence of collective spin states until an equilibrium will be attained dependent on the initial excitation. This may take hours or even days. See, for instance, "Neural networks and physical systems with emergent collective computational abilities", J. J. Hopfield, Proc. Natl. Acad. Sci. USA, Vol. 79, pp. 2554–2558, April 1982, and "Der ganz andere Computer: Denken nach Menschen Art", W. Kinzel et al., Bild der Wissenschaft 1-1988, pp. 37–47. Spin glass systems occur in materials such as Au(Fe) and Cu(Mn) alloys and Cd1-xMnxTe semi-magnetic semiconductors, in electronic dipole glasses (K0.8Na0.2TaO3), electron glasses (compensated Si or GaAs), and vortex glasses in superconductor junction arrays. Apart from the impractically long relaxation times, contacting and learning at such nanoscale levels are problems virtually impossible to solve, at least with today's technology. In addition, full connectivity, which is assumed in the glassy system models, does not really happen in these physical systems due to the finite range of physical interactions.

BRIEF DESCRIPTION OF THE DRAWING

The invention is explained in detail hereinafter by way of example and with reference to the accompanying drawing, wherein:

FIG. 1 illustrates the architecture of a conventional layered neural net;

FIG. 2 gives an example of the neural net of the invention with an inhomogeneous non-linear electrically conducting layer;

FIGS. 3 and 4 illustrate stimulus-response diagrams for the embodiment in FIG. 2;

FIG. 5 shows the functional distribution of inhomogeneities in the layer of FIG. 2;

FIGS. 6 and 7 give background information to formulae in the text relating to the operation of the non-linear inhomogeneous layer;

FIGS. 8, 9, 10 and 11 illustrate a simple quantitative example of the operation of a neural net in the invention;

FIGS. 12, 13, 14, 15 and 16 give detailed examples of the layered structure in a neural net of the invention.

DETAILED DESCRIPTION OF THE EMBODIMENTS

Despite the large variation in their architectures, the basic operation performed by neural nets is a non-linear mapping of the input onto the output in a parallel and, preferably, in an adaptive fashion. Below it is illustrated how an inhomogeneous physical substance with a non-linear stimulus-response characteristic can in principle be employed to represent a neural network. An electronic example of such a net of the invention is studied thereafter. Note that the invention is essentially technology-independent and that the principles of the invention are not restricted to electronic embodiments.

Prior Art

Neural networks are made of a collection of highly interconnected simple processors which, in the most interesting version program themselves adaptively in order to solve a given problem. What really makes neural networks systems rich is their highly varied interconnection styles between the processors. These styles determine the architecture of a neural net which can vary anywhere from a layer-to-layer interconnected feedforward network to a fully interconnected Hopfield network. Despite the large variation in their architectures, the basic operation performed by neural networks still is an adaptive non-linear mapping of the input onto the output in a parallel fashion. This non-linear mapping is demonstrated in a simplified three layer network shown in FIG. 1.

FIG. 1 shows an example of a conventional architecture of a neural net 10. Neural net 10 is comprised of a plurality of successively interconnected neuron layers 12, 14 and 16 between an input 18 and an output 20. Each interconnection (synapse), e.g., 22 weights the signal provided by a source neuron, e.g., 24 by a particular factor (synaptic coefficient) and supplies the weighted signal to the input of a destination neuron, e.g., 26 in the next layer. Destination neuron 26 sums the thus scaled contributions of all its source neurons and applies a non-linear function (e.g., a sigmoid function) to the resulting sum for generating its output signal that in turn is supplied to a neuron, e.g., 28 of the next successive layer via synapse 30. Note that the essential operation of neural net 10 is a functionally distributed and parallel non-linear mapping of the input onto the output.

Principle of Fully Distributed Embodiment

The invention provides a device, which will achieve non-linear mapping like neural nets such as the one of FIG. 1, but replaces the lumped pans of a neural network by fully and substantially continuously, spatially distributed functionalities.

FIG. 2 illustrates an example of a neural net 100 in the invention. Neural net 100 includes a non-linear multi-port inhomogeneous electronic layer 102 and a plurality of input contacts 104 and a plurality of output contacts 106. Input information to layer 102 is supplied through a set of N switches 108 connected to contacts 104. The binary (on or off) state of a particular switch "i" among switches 108 is denoted by X_i, for i=1,2, . . . , N. The states X_i determine which ones of contacts 104 are connected to power supply 110 that provides a bias voltage V. The raw output from layer 102 is assumed to be a set of response currents, I_j for j=1,2, . . . , M, which flow out of the device through small resistors 112 to ground. Although the inputs and outputs can also be purely capacitive, this example considers the response of net 100 to a dc excitation through ohmic contacts 104 and 106. Contacts 104 and 106 are shown to be arranged at the periphery of layer 102, although this does not have to be so in practical implementations. The embodiment of neural net 100 is assumed to be a planar device in this example to be able to facilitate the implementation. The type of vector mapping discussed below emphasizes binary input vector components rather than continuous input signals. This is done to simplify the analysis of the examples.

Assume that an input vector $\underline{X}^m$ is presented as a series of the binary conditions (open=0; closed=1) of switches 108. Superscript "m" differentiates vectors corresponding to different sets of binary input signals. The associated responses, i.e., the components of current vector $\underline{I}^m$, show non-linear correlations between the input vector components X^m_i and the contents of network 100. The contents of network 100 is represented by the distribution pattern of inhomogeneities and non-linearities 114 serving as neural "weights". The general form of response current $\underline{I}^m$ can be written as:

$$\underline{I}^m = V\,G^m\underline{X}^m; \tag{i}$$

where G^m is the non-linear conductance tensor of the network, depending on the pattern m and on bias voltage V. Note the structure of formula (i) reflecting that of Ohms' law. Written out in its indexed components, formula (i) equals:

$$I^m_j = V\Sigma^M_{i=1}G^m_{ji}(V)X^m_i; \tag{ii}$$

The system under study is assumed not to show negative differential resistance (NDR) properties for this moment. As is discussed below, if layer 102 has NDR, many of the encoding techniques mentioned here get simplified.

Signal Diagrams

Since NDR effects are assumed not to be involved, the expected response currents I^m_j as a function of input bias level V and input vector components X^m_i can be schematically given as shown in FIGS. 3 and 4. FIG. 3 gives the dependence on bias voltage V of current I^m_k, produced at an output contact "k" in response to input pattern $\underline{X}^m$. As is clear, current I^m_k is a monotonic function of V. FIG. 4 shows the dependence of current I^m_k on the total number "n" of closed switches 108. Again, current I^m_k is a monotonic function of additions to the sum $K=\Sigma^M_{i=1}X^m_i$, i.e., of the total number of closed ones of switches 108. Note that the schematic response of any of output contacts 106 as a function of $\underline{X}^m$ shown in FIG. 4, is similar to the response obtained from a functional MOS transistor which couples several capacitive (gate) inputs to a single output current (source-drain) to achieve a neuron-MOSFET. See: "A Functional MOS Transistor Featuring Gate-Level Weighted Sum and Threshold Operations", T. Shibata et al., IEEE Trans. Electron Devices, vol. 39, 1992, pp. 1444–1455.

If, due to the monotonic output response of non-linear layer 102, the output signal were to be defined simply as:

$$O^m = I^m - I_{threshold}; \tag{iii}$$

then closing an additional one of the switches 108 and/or increasing the bias level to V+dV would always give rise to a change of fixed polarity in the output signal O, that is, an increase in above example. In other words, encoding the output information according to (iii) does not provide inhibitory behavior. If non-linear mapping, as in neural nets, is to be achieved, inverse effects such as negative contributions or inhibition should be produced by some of the input signals X^m_i. In addition, it should be shown that electronic signal crossing between two opposite cross coupled pairs of input and output contacts is possible to accomplish signal correlation. To this end, the non-linearities in the system could be used in encoding the information at the output. The non-linearities can have positive or negative sign as opposed to unipolar nature of the net (total) response from our system. Another useful type of encoding of the responses from the non-linear conducting layer **102** could be obtained by enhancing high order effects. This enhancement can be achieved by utilizing differential output signals or relative output signals. For example, the following encoding schemes can be used:

$$O^m_j=I^m_k-I^m_p; \qquad \text{(iv)}$$

$$O^m_j=I^m_j/I^m_{tot}; \qquad \text{(v)}$$

$$O^m_j=\delta_v(I^m_j); \qquad \text{(vi)}$$

The first option in (iv) defines component j of output signal O^m as the difference between the response currents supplied at an output with index "k" and at an output with index "p". The second option in (v) gives the output component j of O^m as the current I^m_j normalized to the total current associated with on/off switch pattern "m" The third option of (vi) is the differential output as a function of changes in bias V, that in itself can be considered as an input signal affecting the response currents as do the input signals X^m_i. The bias voltage change can be, e.g. a ramped, a sinusoidal or a pulsed excitation. Related to this latter case, one can also encode the information in the amplitude of different signal harmonics of the bias voltage change. Note that output functions (iv)-(vi), and especially (iv) and (v), are rather simple and can be implemented easily at the output. A threshold current $I_{treshold}$ can be subtracted from any of the currents I^m_j for detecting the polarity of the results in each case. With these definitions, the encoding described above is analogous to the function of the output neurons in conventional Neural Networks.

It is shown below, with simple examples, that the types of encoding (iv)-(vi) allow for effects like inhibition and signal crossing. To this end, equation (ii) is rewritten as a series expansion:

$$I^m_j=\Sigma^M_{i=1}[G_{1,ji}X^m_iV+[\Sigma^M_{k=1}G_{2,jik}X^m_iX^m_kV^2++\Sigma^M_{p=1}G^m_{3,jikp}X^m_iX^m_kX^m_pV^3+\ldots]]]. \qquad \text{(vii)}$$

To be able to proceed further analytically, consider a non-linear conducting plane **102** with only two input contacts **104** and two output contacts **106**. Then, index m takes the values=0, 1, 2 and 3 for input vectors $\underline{X}^m$ (0,0), (1,0), (0,1) and (1,1), respectively. In addition, assume that the non-linearity in the system is limited to second order terms only. In this case, the currents of (vii) can be written as:

$$I^m_j=\Sigma^2_{i=1}[a_{ji}X^m_iV+[\Sigma^2_{k=1}c_{jik}X^m_iX^m_kV^2]]; \qquad \text{(Viii)}$$

Coefficients "a" and "c" used here replace the linear and non-linear conductances G_1 and G_2, respectively. Equation (viii) represents a set of eight equations, second-order in V, for the two currents. Using equation (viii) and the encoding options of (iv)-(vi), it is shown below how to obtain inhibitory effects and signal crossing in the simplified non-linear conducting layer network of the invention with two input and two output contacts.

For the first encoding option (iv) there is only one real output to consider because, in this case, O^m_1 uses the difference between the currents at the two output contacts. With above formulation, it is easy to show that the output signal of the network for $\underline{X}^3$=(1,1) is:

$$O^3_1=O^1_1+O^2_1+2(c_{112}-c_{212})V^2; \qquad \text{(ix)}$$

Equation (ix) shows that, due to the last term which can be negative, O^3_1 can become smaller than $(O^1_1+O^2_1)$ and, with proper choice of conductivities c_{112} and c_{212}, even negative. With such an output an XOR operation can be realized in net **100** without any negative polarity in the individual currents I^m_j. Since conventional neural nets require negative connections or inhibition to achieve an XOR operation, the result given above shows that the relative responses of non-linear conducting layer **102** also carries inhibitory information if the non-linearities are chosen appropriately. Note that such a result would not be possible with a purely linear layer, because the last term in (ix) would then be missing.

In order to show signal crossing, i.e. a first and a second output signal being each composed of contributions produced by more than one input signal, at least two input and two output contacts **104** and **106** are needed. For a simplified 2-input/2-output conducting layer **102**, the encoding option of (v) is chosen. Using above notation where m=0, 1, 2 and 3 for input vectors $\underline{X}^m$ (0,0), (1,0), (0,1) and (1,1), respectively, and a little algebra, it becomes clear that, in order to have signal crossing, it needs to be proven that the following states are attainable:

$$(O_{11}-O^2_1)<0 \text{ and } (O^1_2-O^2_2)>0; \qquad \text{(x)}$$

or, in plain language, that the contribution of the input at contact "i=2" dominates the output at contact "k=1" and that the contribution of the input at contact "i=1" dominates the output at "k=2". Proving the validity of conditions (x) is equivalent to showing that the difference:

$$\Delta=I^2_2I^1_1-I^1_2I^2_1. \qquad \text{(xi)}$$

can be negative. For a linear conducting layer **102**, this condition would translate into $a_{11}a_{22}<a_{12}a_{21}$, which is physically unacceptable as a larger current would result from the activated input contact farthest away. In order to show that Δ can plausibly be negative with a non-linear conducting layer, assume that all linear elements of the linear conductance matrix are equal and that the non-linearities can be treated as small perturbations. With these conditions, Δ<0 translates into

$$C_{122}+C_{211}>C_{111}+C_{222}; \qquad \text{(xii)}$$

This last condition can be satisfied easily by sub-linear (saturating) direct conductances and superlinear cross conductances. It is very easy to demonstrate this effect in a "Y" resistor network where each branch contains a linear resistor except one branch that includes a non-linear resistor. If: a) one of the outputs is connected to the non-linear resistor and the other output to a linear one; and b) one of the inputs is connected to the remaining linear resistor and the other to the centre node, then application of (1,0) and (0,1) inputs can generate differential (0,1) and (1,0) outputs with proper choice of resistor values and non-linearity. Similar examples can be given for the encoding technique stated in (vi). As stated earlier, in this last case, also output harmonics of periodic input signals can be utilized in addition to slopes of conductance changes.

Physical Operation

In order to go into more detail, consider a device with a two-dimensional non-linear conducting layer representing an inhomogeneous surface inversion layer in a field effect device. It is assumed that this layer is composed of an array of electron puddles with mutually different electron densities due to the differences in the surface potential distribution. It is shown below that the intrinsic non-linear vector mapping abilities of a non-linear surface conducting layer in this simplest form are attained. Any additional features introduced at a later stage will then enhance the effects. The physics of the device concept studied here may be considered to be similar in nature to the experimental MOS-device of Shibata cited above. However, the prior art device of Shibata only serves to implement the synaptic connections for a single Neuron. In sharp contrast to the cited art, the non-linear electronic layer in the invention is utilized as a distributed neural transformation system to act as a full neural network between multiple input and multiple output terminals. This crucial difference may require proper encoding as described in the previous section to be able to extract the required information.

FIG. 5 schematically illustrates such a device 300 in the invention. Device 300 functionally comprises an inhomogeneous array of dots representing locations of free electron puddles such as 302 and 304 in an inhomogeneous surface inversion layer in a field effect device. Puddles 302 and 304 are formed by trapped positive charges within an insulator (not shown) at the surface. These electron puddles are electrically coupled to one another, here schematically illustrated only for the pair of puddles 302 and 304, by means of non-linear capacitances, e.g., 306, and non-linear conductances, e.g., 308. Each respective one of the puddles further is capacitively coupled to ground, indicated here for a single puddle 310 by way of capacitance 312. The input/output signals are provided at contacts on the periphery via linear resistors, of which only resistors 314 and 316 are shown. As indicated above, the input/output contacts do not always have to be at the periphery, but can also be distributed within the surface area. This latter option may be especially beneficial in, for example, image processing applications where the input is supplied optically to the frontal surface as a two-dimensional signal.

The transport in the layer shown in FIG. 5 is modeled by simplified and normalized equations approximating the basic features of the conduction in an inhomogeneous surface inversion layer of a large area sub-threshold semiconductor device. For background information on the physical aspects, see Shibata. The non-linear current flowing among the puddles is given by:

$$I_{dot,i}=\exp(\alpha/T) \text{ for } \alpha<0;$$

$$I_{dot,i}=1+(\alpha/T) \text{ for } \alpha>0; \qquad \text{(xiii)}$$

wherein $I_{dot,i}$ represents the net electron current flowing out of a puddle, the puddle being denoted by index "i", wherein α is defined by

$$\alpha=(E_{F,i}-E_{bar,i}); \qquad \text{(xiv)}$$

$E_{F,i}$ being the Fermi level associated with puddle "i", $E_{bar,i}$ representing the height of the potential barrier between puddle "i" and one of its neighbours "j", and wherein T represents a normalized absolute temperature. Barrier height $E_{bar,i}$ in a real device depends on the potential appearing between electron puddles and is given by the following empirical relation:

$$E_{bar,i}=E_{c,i}+|b_0-b_1(E_{c,i+1}-E_{c,i})|; \qquad \text{(xv)}$$

wherein $E_{c,i}$ is the conduction band edge and b_0 and b_1 are constants. For a representation of these quantities see FIG. 6.

The capacitances between two puddles are assumed to depend on the number of carriers contained within these puddles and they are given by:

$$C_{ij}=d_0/(d_1-q_i-q_j); \qquad \text{(xvi)}$$

wherein d_0 and d_1 are constants and q_i and q_j represent the number of charges in puddle "i" and neighbouring puddle "j". This equation simply approximates the fact that capacitances, e.g., 306 between puddles containing fewer charges are smaller. The capacitance 312 between each puddle and ground is taken as a constant. No special attention has been given on fitting the form or the parameters appearing in these equations to real devices. For this, one has to start with device level equations treating (in this example) MOS physics properly. This is not necessary for the present, as only the basic concept of vector mapping in a general non-linear layer is presented. The exact form of the non-linearities appears to be not crucial to the operation of the conducting layer as a vector mapping network. The response of the model system described above is calculated numerically by using the discretizised Poisson equation and the current continuity equation.

With the model described above, first the input bias dependence of the conductances, e.g., 308 in the layer is examined. This is done by applying the input bias voltage V at only one of the inputs of the device and observing the total current across the device as a function of the value V of the input bias. The results are shown in FIG. 7 for two different values of the input/output resistances 314 and 316. The basic appearance of these characteristics is similar to the subthreshold response of a single MOSFET. In the present case however, we functionally do have a collection of interconnected floating sub-threshold MOSFETs. The results in this figure can also be interpreted as a bias-dependent "percolation" through the puddles of the device. There is no sharp cut-off to zero current at low biases, in contrast to real percolative models, due to the nonzero temperature parameter taken into account in the calculations.

Next, the input/output relation is discussed for a 3-input/5-output example 400 of device 300 with reference to FIGS. 8, 9 and 10. Device 400 includes a non-linear conducting layer 402 with three input contacts 404, 406 and 408, that can be connected to power supply V through switches 410, 412 and 414. Layer 402 further includes five output contacts 416, 418, 420, 422 and 424 to supply response currents I1–I5 dependent on which ones of switches 410–414 are closed. The response currents are fed to a circuit 426 to determine differential quantities as discussed under equation (iv) above. Note that circuit 426 can simply be a resistance coupled between two output contacts. The voltage over this resistor then is indicative of the net current through the resistance.

FIG. 9 gives a Table I showing the values of response currents I1–I5 in predetermined units, e.g., mA, as a function of input patterns X of open or closed ones of switches 410–414. In vector notation, $X=(X1, X2, X3)^T$, wherein X1 corresponds with the state of switch 410, X2 corresponds with the state of switch 412 and wherein X3 corresponds with the state of switch 414. Component Xi equals unity if the associated one of switches 410–414 is closed, thereby connecting power supply V to the appropriate one of contacts 404–408, and is zero otherwise. Output contact 422 is assumed to be disconnected for this example. FIG. 10 gives a Table II giving the differential values between currents I2 and I1, and between currents I3 and I5. The information in Table II is visualized in the diagram of FIG. 11. FIG. 11 clearly demonstrates how output signal I5–I3 changes from positive to negative, or, in other words, provides inversion when the system receives more input signals. Output signal I2–I1 on the other hand demonstrates the XOR operation, and therefore signal crossing, as the input vectors (100), (010), (110) and (000) produce outputs of (+), (+), (−), and (−) polarity, respectively, with regard to, e.g., a level of I2–I1 equal to 3.5. A logic output signal is simply produced by way of appropriate thresholding and subsequent limiting. Other encoding schemes than the one providing differential outputs, such as the ones in equations (v) and (vi), give inversion and signal crossing as well. The output examples given here are obtained after conducting non-linear layer 402 has reached a steady equilibrium state. If the output current values are probed before such a steady state is reached, then the relative values of the output currents in such a conducting plane can be totally different from the ones listed in Table II. Intermixing of signals in this transient regime may turn out to be more easily achievable as compared to the intermixing in the steady state situation.

A negative-differential-resistance (NDR) effect is attained in the inhomogeneous inversion layer array by letting the current decrease slightly as a function of increasing bias voltage difference across the electron puddles. Due to this NDR, the relative locations of output current values are highly intermixed for different input vector configurations. In addition, the input node current values also show such non-linear intermixing. It is obvious that such responses result in non-trivial non-linear vector mappings by simple thresholding without the need to apply any of the coding methods given in (iv), (v) and (vi).

FIG. 12 shows a portion of a non-linear inhomogeneous electrically conducting transport structure **500** for use as a layer **102** in a neural net of the invention. Structure **500** includes a semiconductor substrate **502**, for example of P-doped silicon, and an electrically insulating layer **504**, for example silicon dioxide. Layer **504** is provided with charge trapping regions **506**, **508**, **510**, **512**, **514** and **516**, for example of polycrystalline silicon or silicon nitride. Electrical charges of a predetermined polarity type, positive in this example, trapped in one or more of regions **506–516** induce inversion regions, such as **518**, **520**, **522** and **524**, and depletion regions, such as **526** and **528** in substrate **502** near insulating layer **504**. Trapped electrical charges of the other polarity, here negative, induce accumulation regions, such as **530**, in substrate **502** near insulating layer **504**. Thus, the trapped charges specify location-dependent conductivity properties in substrate **502** establishing an electrical transport structure **500** based on field effect operation.

FIG. 13 illustrates a portion **600** of structure **500** in a neural net of the invention. Portion **600** includes pan of substrate **502** and of insulating layer **504** discussed under FIG. 12. Insulating layer **504** houses charge trapping regions **606**, **608**, **610** and **612** in order to create inversion regions, depletion regions and accumulation regions as discussed previously. An ohmic contact **614**, e.g., of aluminum, to an N-doped region **616** in P-doped substrate **502** is used to supply an input signal or to transport the response current effected at that location by the input signals. Alternatively, contact **614** can be used as an additional long range interconnection to galvanically or capacitively couple two or more remote areas of substrate **502**, e.g., in order to circumvent short range effects due to screening. Contacts **618** and **620** are used as capacitive inputs to substrate **502**, either directly such as **614** or indirectly via charge storage regions **608** and **610**. Again, contacts **618** and **620** are used individually to apply an externally generated input signal or an internally generated signal by way of feedback mechanism.

Note that the structure in. FIG. 13 utilizes a fairly homogeneous and substantially uniformly doped substrate **502**. The inhomogeneities induced by the charges trapped in regions **606–610** are only perturbations superposed onto a homogeneous configuration. This simple arrangement suffices to achieve non-linear and fully distributed mapping as in a conventional neural net. However, substrate **502** may in addition contain an array of N-doped regions (not shown here) or heterostructure regions (not shown here) adjacent insulating layer **504** in order to enrich the electric transport properties by providing negative differential resistance or optoelectronic effects.

Long Range

The above mentioned screening effects will occur in systems that are large in relation to the range of certain interactions. These effects will adversely affect the discrimination between responses to different sets of input signals. Therefore, beyond a given size of the system in the invention, resolution and noise become important. One way of compensating these drawbacks is to use more contacts to extract more responses that combined produce the output signal(s). Another way, already briefly mentioned above, is to provide a few long-range interconnects, conductive or capacitive, between various locations distributed over the device. This could be attained by a pattern of substantially equal conductance distance between the locations interconnected. FIG. 14 illustrates an example of such an arrangement **700** that is implemented in a technology similar to that for portion **600** of FIG. 13. Arrangement **700** is comprised of a semiconductor substrate **702** with a plurality of input contacts and output contacts **704** and **706**, a plurality of inhomogeneous domains such as **708** and **710**, and a plurality of long range interconnects such as **712**, **714** and **716**. Long range interconnects **712–716** serve to equalize an avenge electrical path length between different ones of input-output contact pairs. Note that since arrangement **700** is inhomogeneous in large scales as well in a conductive state, it is expected that the number of long range interconnects **712–716** is relatively small. Note that wafer-scale integration is a suitable technology to implement vast versions of such a system.

Learning

In order to provide teaming facilities, some measures are needed to selectively modify the transport properties of substrate **502**, for instance, by modifying the amount of charge trapped in trapping regions **506–516**. As known, conventional neural nets learn by means of adapting the strengths (or weights or synaptic coefficients) of the synapses on the basis of training examples. A well known learning strategy is the backpropagating mechanism applied to a layered net, such as the one of FIG. 1. Therein, a discrepancy between an obtained output signal of layer **16** and a desired output signal derroes an error quantity that is minimized by changing the synaptic coefficients according to a steepest descent algorithm. This method is based on the individuality of each neuron functionality and of each synapse functionality. Note that in the present invention individual neurons and individual syfiapses cannot be distinguished.

FIG. 15 gives an example of a neural net in the invention comprising an electrical transport structure **800** including learning facilities. As structure **800** is partially similar to portion **600**, reference is made to FIG. 13 and its discussion above for the electrical transport aspects. Structure **800** includes a semiconductor substrate **802** of, e.g., P-doped silicon, an electrically insulating layer **804** of, e.g., silicon dioxide, a charge trapping layer **806** of, e.g., polycrystalline silicon, islands of polycrystalline silicon embedded in insulating layer **804**, or silicon nitride, a piezoelectric layer **808** of, e.g., zinc oxide, for the propagation of mechanical stress waves and electrically isolated from layer **806**, an insulating layer **810** of, e.g., silicon dioxide, and an electrically conductive layer **812** such as aluminum or tungsten, functioning as a (global) gate as in a field effect transistor.

Learning is as follows. Assume that the initial trapped charge distribution in layer **806** gives rise to incorrect responses when being given a particular set of predetermined input signals. Then mechanical stress waves are created in piezoelectric layer **808** of finite size. The waves may be stochastic. A standing wave pattern is established due to reflections throughout layer **808**. This in turn produces an standing electric field pattern that affects the electrical transport properties in substrate **802**, in addition to the electric field of the trapped charges in layer **806**. The stress pattern either improves the responses or worsens them. If the stress pattern improves the responses, the perturbation produced by the associated electric field is frozen. The effects of the stress field are replaced by a modification of the trapped charge distribution. One way of doing this is to apply a bias voltage at global gate **812**. Such a bias voltage should be just below the threshold of tunnelling between the charge trapping layer **806** and substrate **802**. This is comparable to the write/era threshold of EEPROMs. The electric field produced by the spatially varying stress field is added to the constant bias of global gate **812**. As a result, the total electric field between charge trapping layer **806** and substram **802** is selectively controlled to locally exceed or stay below the threshold, thereby selectively enabling charge tunnelling. This type of learning may be termed "Stochastic Learning".

Many variations can be included in the application of the bias, stress field and charge trapping. For example, with proper design, charging may be achieved by charge transport to and from global gate **812**. A plurality of mutually electrically isolated global gates **812** may be provided in a predetermined geometrical pattern. Note that, for example, GaAs is piezoelectric. An epitaxial GaAs layer separated from a GaAs substrate by an insulating layer may provide both piezoelectric facilities and electronic signal transport. Also layers **806–810** may be replaced by a layer of ferroelectric material. As known, a ferroele ctric is a material having domains whose electric dipole field can be selectively modified. Ferroelectric materials are used in certain types of non-volatile memories.

FIG. **16** illustrates another example of a layered structure **900** for use as a cell in a learning neural net in the invention, based on opto-ele ctronics. Structure **900** includes a semiconductor substrate **902**, for example of P-doped silicon, an insulating layer **904**, for example of silicon dioxide, and a charge trapping region **906**, for example of polycrystalline silicon, embedded in insulating layer **904**. An upper surface **908** of structure **900** carries electrodes **910** and **912** connected to opposite poles of a power supply (not shown). Electrode **912** is connected to a photodiode **914** and electrode **910** is connected to a resistor **916**. Photodiode **914** and resistor **916** are interconnected via a node **918** located near charge trapping region **906**. Incident light rays **920** on photodiode **914** cause the potential at node **918** to change, thereby locally affecting the charge distribution in, and hence the transport properties of the substrate. In case of an improved output (lag) of the neural net due to these changes, the trapped charge in region **906** is adapted by applying a large bias voltage, e.g., via electrodes **910** and **912**, to cause tunnelling between electrodes **910** or **912** and charged trapped region **906**.

A plurality of such cells of photodiodes and resisters can be arranged in a predetermined configuration, e.g., in a rectangular grid or in a radial fasliion to imitate the retina, in order to enable pattern recognition. Each pixel of an image then is mapped onto a corresponding photodiode thereby affecting the associated transport properties in terms of depletion layers **920** and inversion layers **922**.

I claim:

1. A neural network processor comprising:
input means for receiving a plurality of input signals;
output means for providing at least one output signal;
neural network means coupled to the input means and the output means, the neural network means comprising a physical body, the physical body having physical properties allowing for propagation of an electrical WAVE in an unchanneled fashion therethrough from the input means to the output means, the body comprises:
at least one region, which is suitable for affecting electrical conduction through the physical body;
training means for altering a state of the region during a learning mode of the neural network means; and wherein:
the physical body comprises a semiconductor body; and
the semiconductor body comprises:
a semiconductor substrate;
an insulating layer on a major surface of the substrate;
at least one charge storage region within the insulating layer for inducing tho at least one region in the substrate.

2. The processor of claim **1** wherein the training means comprises at least one conductor situated on the opposite side of the insulating material from the semiconductor substrate.

3. The processor of claim **2** further comprising an additional doped region within the semiconductor substrate, which doped region is in electrical contact with the conductor, via an aperture in the insulating material.

4. The processor of claim **3** wherein the semiconductor substrate has a doping of opposite conducting type to that of additional doped region.

5. The processor of claim **2** wherein the conductor is situated directly over the charge storage region and interacts capacitively or galvanically therewith.

6. The processor of claim **2** wherein the conductor is situated over a thin portion of the insulating material and interacts capacitively or galvanically with the semiconductor substrate.

7. The processor of claim **1** wherein the training means comprises
a piezo-electric layer on the insulating layer for receiving mechanical stress waves which alter the electrical properties of the piezo-electric layer;
a second insulating layer on the piezo-electric layer; and
a conductive layer on the second insulating layer, so that the piezo-electric layer alters an effect of the conductive layer on the at least one region.

8. The system of claim **1**, wherein the training means comprises an upper surface containing photoelectric elements for receiving photons for altering electrical properties of the upper surface.

9. The system of claim **1** wherein the training means comprises a ferroelectric layer on the insulating layer.

* * * * *

US005315152A

United States Patent [19]

McHardy et al.

[11] **Patent Number: 5,315,162**

[45] **Date of Patent: May 24, 1994**

[54] **ELECTROCHEMICAL SYNAPSES FOR ARTIFICIAL NEURAL NETWORKS**

[75] Inventors: **John McHardy**, Westlake Village; **Carl W. Townsend**, Los Angeles; **Lin R. Higley**, Laguna Hills; **Frank A. Ludwig**, Rancho Palos Verdes, all of Calif.

[73] Assignee: **Hughes Aircraft Company**, Los Angeles, Calif.

[21] Appl. No.: **770,817**

[22] Filed: **Oct. 4, 1991**

[51] **Int. Cl.^{5}** **G06G 7/06; H03H 11/00**

[52] **U.S. Cl.** **307/201**; 395/24; 338/80; 338/94

[58] **Field of Search** 307/201; 395/21, 23, 395/24; 338/80, 82, 94

[56] **References Cited**

U.S. PATENT DOCUMENTS

3,222,654	12/1965	Widrow et al.	301/201
3,453,602	7/1969	Stewart	395/24
4,945,257	7/1990	Marrocco, III	307/201
4,949,064	8/1990	Minoura et al.	338/80

Primary Examiner—William L. Sikes
Assistant Examiner—Kenneth Parker
Attorney, Agent, or Firm—M. E. Lachman; M. W. Sales; W. K. Denson-Low

[57] **ABSTRACT**

An electrochemical synapse adapted for use in a neural network which includes an input terminal and an output terminal located at a distance of less than 100 microns from the input terminal. A permanent interconnect having controllable conductivity is located between the two inputs. The conductivity of the permanent interconnect is controlled by either growing or eliminating metallic whiskers between the inputs. The growth and elimination of whiskers provides a rapid and controllable electrochemical synapse. Partial neural network systems are disclosed utilizing the electrochemical synapse.

22 Claims, 1 Drawing Sheet

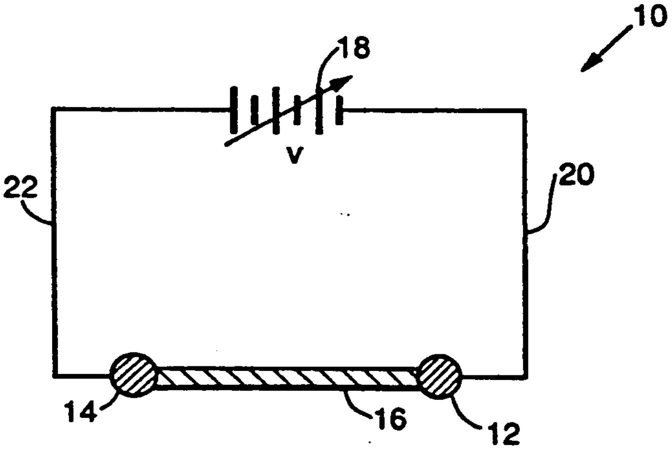
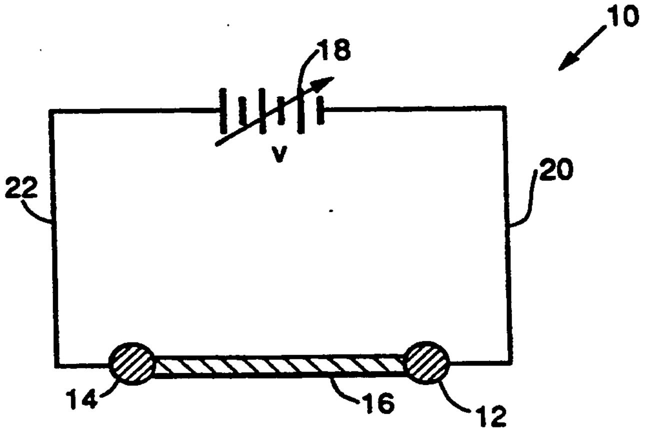


FIG. 1.



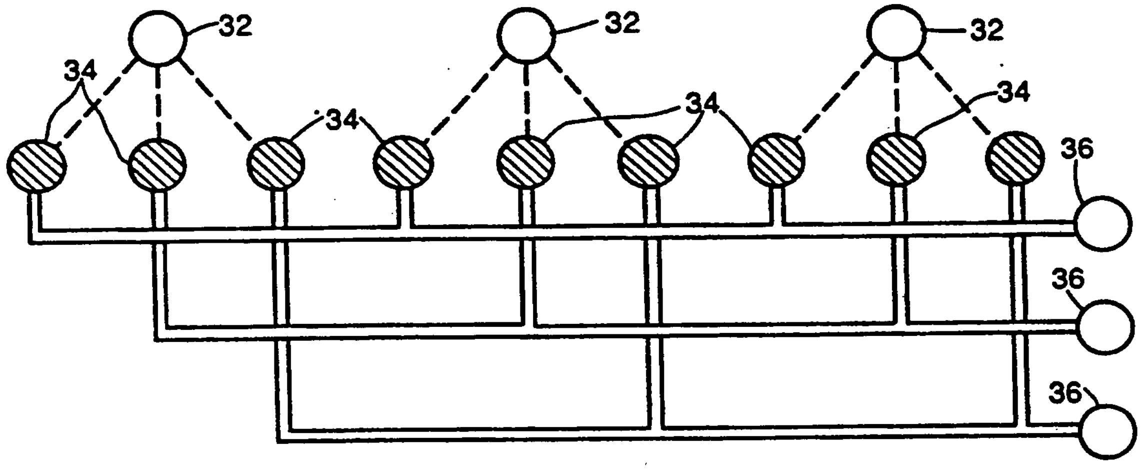
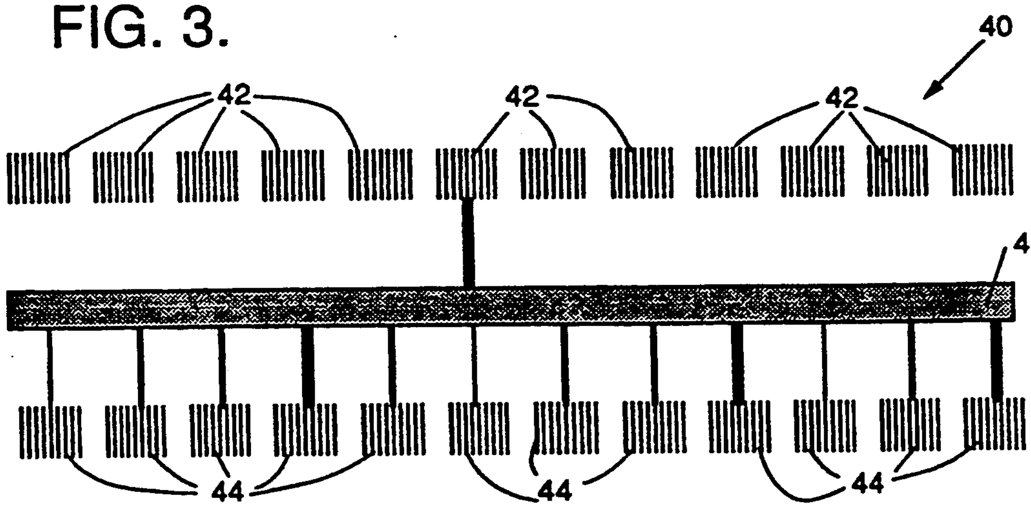


FIG. 2.

FIG. 3.



ELECTROCHEMICAL SYNAPSES FOR ARTIFICIAL NEURAL NETWORKS

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention relates generally to neurons in an artificial neural network. More particularly, the present invention relates to solid state, adjustable weight synapses for controlling the interaction of the neurons in such artificial neural network.

2. Description of Related Art

There has been a great deal of interest in developing non-volatile associative electronic memories based on models of neural networks. A key feature of artificial neural networks is the vast number of synapses that must interconnect each neuron with many others.

Several different approaches have been tried for providing the vast number of synapses required in any artificial neural network. One example is an optical approach, wherein light beams can cross one another without signal interference and, hence, offer a convenient way to form multiple interconnections. However, the lenses and lasers needed for such optical interconnective devices make such systems too bulky for broad applications. (See, for example, J. Kinoshita and N. G. Palevsky, "Computing with Neural Networks," *High Technology*, May, 1987, 24–61, 1987).

An electrochemically regulated synapse known as the "Memistor" was developed in the 1960's by Bernard Widrow, as part of a network known as the "Adaline" network, as disclosed by B. Widrow, in the publication "Neural Network Theory, Past and Present," Paper presented at the DARPA Neural Network Study Symposium, Lincoln Labs, 1987. The Memistor is an electrochemical cell in which copper is either plated on or deplated from a carbon rod. As a result of the controlled plating and deplating of copper, the resistance of the rod is continuously adjustable from 1-10 ohms. This provides a 10:1 range of synaptic "weights." The Memistor serves well from the standpoint of trainability, surviving numerous plating and deplating cycles. However, the Memistor does not lend itself to miniaturization and the device is not practical for large-scale networks.

Metal migration is an electrochemical process related to electroplating. Metal migration takes place between conductors in an active electronic circuit in the presence of a moisture film. Under the influence of a DC voltage, metal ions dissolve from the positive conductor (the anode). The dissolved ions migrate through the moisture film (the electrolyte) and plate out on the negative conductor (the cathode). The deposit often takes the form of metallic whiskers which eventually reach the anode and create an ohmic contact.

Metal migration has been observed with all of the metals commonly used in the electronics industry, but it occurs most readily with silver (see A. Dermarderosian, "The Electrochemical Migration of Metals," Proc. 1978 *Microelectronics Symp.*, 134–141, International Soc. for Hybrid Microelectronics, 1978). The minimum or "critical" voltage V_c required to grow metallic whiskers can range from a few millivolts to over 2 volts, depending on the metal and prevailing conditions surrounding the electronic circuit. Once V_c is exceeded, growth rates tend to increase linearly with $(V-V_c)$ (see P. B. Price, et al., "On the Growth Properties of Electrolytic Whiskers," ACTA Met., 6, 1968). The initial contact resistance is typically in the range of 10^4–10^6 ohms, but with continued whisker growth, the contact resistance falls several orders of magnitude.

It has been observed that, depending on the medium, whisker growth between copper conductors proceeds either from the cathode to the anode (as usual), or from the anode to cathode, which is the reverse of other metals (see A. Dermarderosian, "Raw Material Evaluation through Moisture Resistance Testing," Proc. IPC, 1976). When growth proceeds from anode to cathode, the whiskers are known as "Conductive Anodic Filaments" or CAF (see J. P. Mitchell and T. L. Welsher, "Conductive anodic filament growth in printed circuit materials," Proc. Printed Circuit World Convention II, Session 2A, pp. 42–55 (1981); J. N. Lahti, R. H. Delaney, and J. N. Hines, "The characteristic wearout process in epoxy-glass printed circuits for high density electronic packaging," Proc. *Reliability Phys. Symposium*, San Francisco (1979); J. Lando, J. P. Mitchell, and T. L. Welsher, "Conductive anodic filaments in reinforced polymeric dielectrics; formation and prevention," ibid. (1979); T. L. Welsher, J. P. Mitchell and D. J. Lando, "CAF in composite printed circuit substrates: characterization, modeling, and a resistant material", *Annual Report*, Conference on Electrical Insulation and Dielectric Phenomena, 234–239 (1980)).

It has been suggested that the undesirable growth of metal whiskers observed in electronic circuits be utilized positively to provide the weighted or adjustable synapses in a neural network. Attempts to utilize metal whiskers as interconnects or neural network synapses were not successful, however, because of the inability to reverse the process of whisker growth. In order for metal whisker growth to provide practical synapse interconnects, it is crucial that the whisker growth process be reversible.

In view of the above, it would be desirable to provide electrochemically regulated synapses adapted for use in neural networks wherein the synapse is a solid state configuration which is adapted for miniaturization for use in large scale neural networks.

SUMMARY OF THE INVENTION

In accordance with the present invention, electrochemical synapses are provided for use in neural networks wherein the electrochemical synapses are solid-state devices which are well suited for miniaturization to provide the extremely large number of synapses which are required for a neural network. The invention is based upon providing a permanent interconnect between the input and output terminals of a synapse, such that the conductivity of the synapse is regulated by means of solid-state electrochemistry.

As one feature of the present invention, the conductivity of the permanent interconnect of the synapse is regulated by increasing or decreasing the chemical reactivity (e.g., pH) of the moisture film absorbed on the permanent interconnect. Variations in pH or other chemical properties result in variations in whisker growth and provide a way to actually reverse metallic whisker growth. The ability to vary and even reverse whisker growth by controlling chemical properties provides a solid-state, electrochemical synapse which is especially well suited for use in neural networks.

As another feature of the present invention, the growth of metallic whiskers between the input and output terminal of the synapse is controlled by varying the DC voltage applied across the permanent interconnect. Permanent interconnects made from carbon having an absorbed moisture film present thereon are well suited for producing permanent interconnects on which metallic whisker growth can be controlled in accordance with the present invention by pH and DC voltage changes.

As another feature of the present invention, permanent interconnects are provided which include mixed halides of rubidium with copper or silver. Such interconnects are solid-state systems which have conductivities that can be reversibly controlled. In addition, ion insertion compounds also provide suitable permanent interconnects having conductivities which can be regulated in accordance with the present invention.

The solid state electrochemical synapses, in accordance with the present invention, are well suited for use in neural networks where the synapse terminals are spaced apart at distances of less than 100 microns. The rate at which metallic whiskers can be grown or eliminated is such that conductivity over such interterminal distances can be rapidly varied to provide the rapid changes in synapse memory or weighting required for neural networks.

The above discussed and many other features and attendant advantages of the present invention will become better understood by reference to the following detailed description when taken in conjunction with the following drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a schematic representation of a single electrochemical synapse in accordance with the present invention.

FIG. 2 is a schematic representation of a simple interconnect arrangement which includes a plurality of electrochemical synapses in accordance with the present invention.

FIG. 3 is a schematic representation of a multilayer interconnect arrangement including a plurality of synapses in accordance with the present invention.

DETAILED DESCRIPTION OF INVENTION

The present invention involves solid-state, electrochemical synapses which are adapted for use in neural networks. A preferred exemplary electrochemical synapse in accordance with the present invention is shown generally at **10** in FIG. **1**. The electrochemical synapse includes an input terminal **12**, an output terminal **14**, and a permanent interconnect **16** located therebetween. The permanent interconnect **16** forms an electrolytic path between the input terminal **12** and output terminal **14**. The permanent interconnect has a small, but finite conductivity. The input terminal **12** and output terminal **14** are spaced apart a distance of less than 100 microns. Preferably, the spacing between the input terminal **12** and output terminal **14** will be on the order of 5–10 microns. A DC voltage is provided across the permanent interconnect **16** by voltage source **18** which is connected to the input terminal **12** and output terminal **14** by way of electrical connections **20** and **22**, respectively.

The input and output terminals **12** and **14** are made from suitable metals, such as gold, indium, palladium, platinum, bismuth, silver, cadmium, tin, copper, or lead. As will be discussed below, in one embodiment of the present invention, it is required that one of the terminals be a non-migratable metal such as gold, indium, palladium or platinum, and that the other terminal be a migratable metal such as copper, bismuth, silver, cadmium, tin, or lead. The permanent interconnect **16** is made from a variety of different materials, depending upon the particular means which is being used for controllably varying the conductivity of the interconnect **16**.

A preferred embodiment in accordance with the present invention involves controllably growing metal whiskers between the terminals in order to provide variable conductivity. A carbon channel or other carbon deposited layer is the preferred permanent interconnect for the growth of metal whiskers. The absorbed moisture present in the carbon-based permanent interconnect provides an electrolytic medium in which metal whiskers may be grown. Although a number of different metal whiskers may be grown in accordance with the present invention, copper and silver whiskers are preferred.

As previously mentioned, in certain media, the growth of whiskers between copper terminals, can proceed in a direction from anode to cathode, which is the reverse of other metals. The copper whiskers are then known as conductive anodic filaments or CAF. Their growth takes place by anodic dissolution of the positively biased conductor, migration of the dissolved copper ions down the voltage gradient and precipitation of the copper in the form of a conductive copper oxide.

The precipitation of copper occurs because of pH changes associated with parallel electrode reactions involving the water present in the electrolytic solution provided by the absorbed moisture. Under the influence of an applied voltage, the water reacts at the anode to yield oxygen and hydrogen ions (acid) and at the cathode to yield hydrogen gas and hydroxyl ions (base).

The solubility of copper ions decreases as the pH rises, so that they remain in solution only in a narrow zone close to the anode. As the copper ions migrate into the more neutral electrolyte displaced away from the anode, the copper ions precipitate as the low-density oxide filaments. The spongy oxide product fills the narrow gap to the anode so that the filaments appear to grow directly from the anode.

The copper oxide whiskers grow preferentially along pre-existing paths. This preferred growth path is believed to be due to the hygroscopic nature of copper oxide which would tend to enhance the absorption of moisture. Once a copper oxide whisker connects the two electrodes, the resistance falls progressively with time. This fact, coupled with the controlling influence of pH, provides a capability for controlling whisker growth and removal.

Changes in pH are particularly effective for controlling the growth of CAF. CAF grows by anodic dissolution of copper and precipitation of copper oxide at high pH. By using a reverse DC bias, in which the anode is the non-migratable metal, no metal dissolves and a low pH is produced by the only other possible reaction, the oxidation of water. The resulting acidity redissolves the CAF. Experiments have shown that new whiskers tend to grow most readily in the presence of already established whiskers. This result indicates a mechanism of parallel electronic and ionic pathways. Therefore, the ionic pathway remains operative in the presence of the electronic whisker short and can lead to the gradual removal of that short by acid dissolution.

Deplating can be carried out in a similar manner. For metals such as silver, cadmium, tin and lead, the reverse DC bias dissolves metal from the adjacent whisker metal, but not from the non-migratable metal terminal. The result is to replate the whisker metal onto the original contact or terminal. The reverse DC bias is applied either to specific synaptic connections or it may be applied as a low-level back bias to all connections in a system to provide a form of controlled "forgetfulness".

In another preferred embodiment, the carbon permanent interconnect is replaced with a solid electrolyte such as the mixed halides of rubidium with copper or silver. (See W. Van Gool, Solid Electrolytes, Academic Press, New York, 1978). Whisker growth then entails moving ions in and out of the substrate with no moisture film being required.

Solid electrolytes are part of a large class of materials known as ion insertion compounds such as those described by D. F. Shriver and G. C. Farrington in "Solid Ionic Conductors", Chem. and Eng. News, May 20, 42–57 (1985). In general, the compounds have both ionic and electronic conductivity. Use of these ion insertion compounds provides the ability to control conductivity at both a zero level and at much higher levels utilizing the insertion compounds as a source of ions.

A particularly preferred ion insertion compound for use as the permanent interconnect is tungsten trioxide. Tungsten trioxide is a poor conductor when pure, but in the presence of moisture and DC voltage, it becomes electrochemically doped with hydrogen ions. This results in the formation of a metallic conductor known as hydrogen tungsten bronze (See P. J. Shaver, Appl. Phys. Lett., 11, 255, 1967.)

The use of an ion insertion compound, such as tungsten trioxide, allows the required reversal and control of metal deposition by the process of corrosion. Specifically, electrons released by metal ionization are scavenged by oxygen in the environment surrounding the interconnect. In the case of hydrogen tungsten bronze, a gradual reversion to tungsten trioxide will occur. The electrochemical reaction is:

$$xH_2O+WO_3+Xe^-=H_xWO_3+xOH^-$$

The balancing reaction at the other electrode is:

$$2H_2O=O_2+4H^++4e^-$$

The above two reactions proceed from left to right during hydrogen entry and from right to left during hydrogen loss. The slow reaction of hydrogen tungsten bronze with oxygen is one example of how manipulation of the environment can be used to gradually weaken a synapse to provide a controlled rate of conductivity. General exposure to oxygen, therefore, causes a homogenous conversion of H_xWO_3 to WO_3 as an alternative to the rapid (approximately one second) electrochemical reversal process.

Exposure to oxygen can also be used to dissolve metallic whiskers in the presence of hydrogen ions e.g.:

$$2H^++1/2O_2+M\rightarrow M^2+H_2O$$

where M represents any non-noble metal, such as copper, nickel, iron, zinc, or cadmium.

There are several other examples in which the conductivity of the interconnect is regulated by means of reactive gases. One example involves the use of an acidic gas such as carbon dioxide to lower the pH of the adsorbed moisture film or electrolyte. The acidity tends to redissolve the material. This leads to reversal of whisker growth and resulting reductions in conductivity. Carbon dioxide is particularly preferred for copper based systems because it can also act as a mild complexing agent to enhance metal solubility. Other reactive gases which may be used to control growth and dissolution of the metal whisker include chlorine, bromine vapor, iodine vapor, sulfur dioxide, hydrogen chloride, hydrogen bromide, hydrogen iodide, nitrogen oxides, ammonia, carbon monoxide, or hydrogen.

FIGS. 2 and 3 diagrammatically depict possible multiple interconnect systems utilizing electrochemical synapses in accordance with the present invention. FIG. 2 illustrates a two dimensional multiple interconnect system and FIG. 3 depicts a three-dimensional interconnect system.

The arrangements illustrated in FIGS. 1–3 may be implemented using any state-of-the-art printed wiring board (PWB), integrated circuit, or hybrid circuit technologies. For example, the two-dimensional arrangement shown in FIG. 2 may be constructed using 2-sided PWB technology. Three options for constructing the more complex arrangement shown in FIG. 3 are: multilayer thick-film technology (using screened conductor pastes and glass-refractory dielectrics), co-fired multilayer ceramic technology, and multilayer thin-film technology (using metal film conductors and polymer dielectrics).

Referring to FIG. 2, a two-dimensional design is shown generally at 30. Input terminals made of gold are shown at 32. The input terminals 32 are placed together with bridge terminals 34 made of silver on one side of the PWB. Output terminals 36, made from gold, are placed on the other side of the PWB. Appropriate traces are shown diagrammatically connecting the gold outputs 36 to the silver bridge terminals 34. The connections between the output traces 36 and bridges 34 are preferably made with plated through-holes. Utilizing the conductivity control methods set forth with respect to FIG. 1, whisker growth and removal are carried out along the paths indicated in FIG. 2 by the dotted lines. The resistance of each path is regulated independently. The result is a combination of electrochemical synapses which can be used to provide a neural network.

A three-dimensional electrochemical synapse system is shown diagrammatically in FIG. 3 at 40. The system 40 utilizes multilayer technology wherein inputs 42 and outputs 44 are laid down as parallel traces on the first and third layers of a multilayer structure. The second layer 46 consists of migratable bridge traces laid down at right angles to inputs 42 and outputs 44. Paths between the three sets of traces are defined by forming vertical vias at the crossing points and filling them with an ion insertion material as previously described. Interconnections are then grown and removed by the techniques previously described.

The black heavy lines interconnecting the inputs 42, bridge 46 and outputs 44 in FIG. 3 illustrates how interconnections are established for one exemplary bridge. The thickness of the black interconnecting lines depicts different levels of conductivity as regulated in accordance with the present invention. Preferably, there are enough bridge traces to allow each input 42 to be connected with every output 44 and vice versa.

FIGS. 2 and 3 are representative portions of large neural networks having many synaptic junctions. The operation of such large-scale neural networks requires short cycle times for training the synaptic junctions. The exact rates of whisker growth and removal in accordance with the present invention depend on the factors previously described.

Preferably, the whisker growth and removal rate will allow growth and removal cycle times of a few seconds. For example, the growth or removal rate of a metal whisker is proportional to the current density in amperes per square centimeter at the whisker tip. Although the total current may be minuscule, the small diameter of a whisker, typically a fraction of a micron, yields current densities of ten amps per square centimeter or more. For a metal valence z, and atomic weight W and a density of ρ grams per cubic centimeter, a current density of 10 amps per square centimeter corresponds to a tip velocity of 10 W/zF ρ cm/sec where F is the Faraday constant.

If silver anodes are used, a tip velocity of 10.6 microns per second is achievable in accordance with the present invention. When nickel is used as a migratable terminal, the growth velocity at the whisker tip is 3.4 microns per second. Growth rates of these magnitudes and even higher are possible in accordance with the present invention.

In large scale neural networks, the interconnection path length will be under 100 microns and preferably no more than a few microns. Accordingly, cycle times for growth and removal of whiskers in accordance with the present invention will be on the order of one second or less.

Having thus described exemplary embodiments of the present invention, it should be noted by those skilled in the art that the within disclosures are exemplary only and that various other alternatives, adaptations and modifications may be made within the scope of the present invention. Accordingly, the present invention is not limited to the specific embodiments as illustrated herein, but is only limited by the following claims.

What is claimed is:

1. An electrochemically regulated interconnection adapted for use in a neural network, said interconnection comprising:

an input terminal;

an output terminal located at a distance of less than 100 microns from said input terminal;

a permanent interconnect between said input and output terminals, said permanent interconnect having a pH, a conductivity and comprising a migratable metal which forms whiskers along said permanent interconnect; and

conductivity control means for regulating the conductivity of said permanent interconnect, said conductivity control means comprising means for increasing or decreasing the pH of said permanent interconnect thereby reversibly controlling the growth of said metallic whiskers.

2. An electrochemically regulated interconnection according to claim 1 which additionally includes means for applying a DC voltage across said permanent interconnect, said conductivity control means comprising means for increasing or decreasing the DC voltage applied across said permanent interconnect.

3. An electrochemically regulated interconnection according to claim 3 further comprising means for reversing the bias of said DC voltage.

4. An electrochemically regulated interconnection according to claim 2 wherein one of said input or output terminals comprises a non-migratable metal and the other of said input or output terminals comprises a migratable metal.

5. An electrochemically regulated interconnection according to claim 4 wherein said non-migratable metal is selected from the group consisting of gold, indium, palladium and platinum and said migratable metal is selected from the group consisting of bismuth, silver, cadmium, tin and lead.

6. An electrochemically regulated interconnection according to claim 1 wherein said permanent interconnect comprises carbon and moisture absorbed on the surface of said carbon.

7. An electrochemically regulated interconnection according to claim 1 wherein said permanent interconnect comprises mixed halides of rubidium with copper or silver.

8. An electrochemically regulated interconnection according to claim 1 wherein said permanent interconnect comprises an ion insertion compound.

9. An electrochemically regulated interconnection according to claim 1 wherein said permanent interconnect comprises carbon and moisture absorbed on the surface of said carbon.

10. An electrochemically regulated interconnection according to claim 9 wherein said input and output terminals comprise copper.

11. A method for operating an electrochemical interconnection for use in a neural network wherein said electrochemical interconnection includes a permanent interconnect between an input terminal and an output terminal, said permanent interconnect having a pH, conductivity and comprising a migratable metal which forms whiskers along said permanent interconnect, said method comprising the step of controllably increasing or decreasing the conductivity of said permanent interconnect by increasing or decreasing the pH of said permanent interconnect to thereby control growth of said metallic whiskers between said input terminal and said output terminal.

12. A method for operating an electrochemical interconnection according to claim 11 wherein said conductivity is controllably increased or decreased by varying a DC voltage applied across said permanent interconnect.

13. A method for operating an electrochemical interconnection according to claim 12 further comprising means for reversing the bias of said DC voltage.

14. A method for operating an electrochemical interconnection according to claim 12 wherein one of said input or output terminals comprises a non-migratable metal and the other of said input or output terminals comprises a migratable metal.

15. A method for operating an electrochemical interconnection according to claim 14 wherein said non-migratable metal is selected from the group consisting of gold, indium, palladium and platinum and said migratable metal is selected from the group consisting of bismuth, silver, cadmium, tin and lead.

16. A method for operating an electrochemical interconnection according to claim 11 wherein said step of controllably increasing or decreasing said conductivity comprises exposing said permanent interconnect to a corrosive gas.

17. A method for operating an electrochemical interconnection according to claim 11 wherein said permanent interconnect comprises carbon and moisture absorbed on the surface of said carbon.

18. A method for operating an electrochemical interconnection according to claim 17 wherein said input and output terminals comprise copper.

19. A method for operating an electrochemical interconnection according to claim 11 wherein said permanent interconnect comprises mixed halides of rubidium with copper or silver.

20. A method for operating an electrochemical interconnection according to claim 11 wherein said permanent interconnect comprises an ion insertion compound.

21. An electrochemically regulated interconnection adapted for use in a neural network, said interconnection comprising:

an input terminal;

an output terminal located at a distance of less than 100 microns from said input terminal wherein one of said input or output terminals comprises a non-migratable metal and the other of said input or output terminals comprises a migratable metal;

a permanent interconnect between said input and output terminals, said interconnect having a conductivity and a pH;

conductivity control means for regulating the conductivity of said interconnect by increasing or decreasing the pH of said interconnect; and

means for applying a DC voltage across said permanent interconnect, said conductivity control means comprising means for increasing or decreasing the DC voltage applied across said permanent interconnect.

22. A method for operating an electrochemical interconnection for use in a neural network wherein said electrochemical interconnection includes a permanent interconnect between an input terminal and an output terminal wherein one of said input or output terminals comprises a non-migratable metal and the other of said input or output terminals comprises a migratable metal and wherein said permanent interconnect has a pH and a conductivity, said method comprising the step of controllably increasing or decreasing the conductivity of said permanent interconnect by increasing or decreasing the pH of said permanent interconnect, said increase or decrease in said conductivity being achieved by varying the DC voltage applied across said interconnect.

* * * * *